A survey of synchronous RAM architectures

Report

Author(s):
Gries, Matthias

Publication date:
1999-04

Permanent link:
https://doi.org/10.3929/ethz-a-004287257

Originally published in:
TIK Report 71

A Survey of Synchronous RAM Architectures

Matthias Gries
Computer Engineering and Networks Laboratory (TIK)
Swiss Federal Institute of Technology Zurich
CH-8092 Zurich, Switzerland
email: gries@tik.ee.ethz.ch

TIK-Report No. 71
April 19, 1999

Abstract

The functionality of volatile random access memories (RAMs) in personal computers, embedded systems, networking devices, and many other products is based on an access scheme which was designed over thirty years ago. Since then a variety of different realizations has evolved. Due to the fact that VLSI designs for memory chips have always been optimized for area and not for access speed, RAM chips have become more and more the performance bottleneck of complex computing systems.

This survey gives an overview of current memory chip architectures. The basic functionality of memories is explained and the advantages and drawbacks of each RAM type are discussed. By providing a better understanding of the limits of current RAM designs, this report supports the decision for a particular RAM in an individual application.

Contents

1 Introduction

The functionality and architecture of static and dynamic random access memories (RAM) were developed at the end of the 60's (static RAM cell functional principle: [67], dynamic RAM cell functional principle: [7]). Astonishingly, the basic functional principles of memory chips have not changed over the years. New semiconductor technologies have evolved and led to a spectacular increase in memory capacity. Since memory chips have always been optimized for capacity and not for access speed, they have become the main performance bottleneck of computing systems, see [3, 75, 4].

Memory chip manufacturers have tried to bypass the weak performance of the memory core by designing tricky memory interfaces which isolate the behavior of the slow memory core from the fast interconnections between the memory chip, a memory controller, and a central processing unit. This has led to a variety of memory interface implementations which essentially are based on the same memory core technology and functionality.

This survey illustrates and compares several currently available RAM implementations. Most of the information has been extracted from data sheets. In addition, performance, capacity, and timing measures of different architectures have been unified. This simplifies the comparison of individual memory chips as well as the choice of a particular RAM which best fulfills the needs of a specific application. The survey does not focus on special electrical characteristics but on performance measures. Although, only single RAM chips are considered, the corresponding module or bus standards are shortly mentioned.

Structure of a minimal memory system

A minimal memory system consists of a central processing unit (CPU), a memory controller, and a RAM chip. The memory controller is responsable for the translation of read and write instructions of the CPU into control signals for the RAM. Linear addresses must be translated into two-dimensional addresses which are used inside the memory chip. In addition, the controller must satisfy the timing requirements of the memory chip because the timing of control signals is not checked by the RAM itself. In the worst-case, the memory controller has to stall the CPU until the RAM is again capable of accepting read or write accesses.

Three buses are distinguished, namely for data, control signals, and addresses. In some systems, combinations of these signals are multiplexed onto a single bus. Usually, this is done with addresses and data signals or addresses and control signals. The signal values on the control bus at a certain point of time can be seen as a memory instruction by which the memory chip is programmed. In this view, suitable control signal combinations define an instruction format. See Fig. 1 for an overview of a minimal memory system.

Related work

A lot of RAM architecture overviews have been published. However, since advances in RAM technology evolve rapidly, only the most recent ones are not outdated.

In [35] and [56], an overview of the functionality, the architecture, and some typical applications of dynamic RAM types such as CDRAM and SDRAM is given. However, some of the described RAM types, e.g. EDO-RAMs, concurrent Rambus RAMs, and Video RAMs, are already outdated. Unfortunately, the articles only show new trends in RAM architecture development. They do not provide a detailed description and comparison of

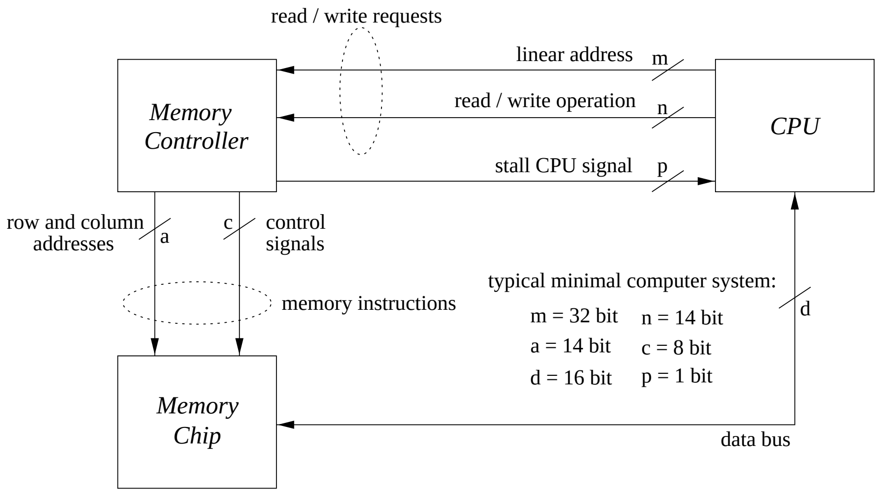

Figure 1: A minimal memory system.

the different memory types in terms of performance measures, electrical characteristics, or organization.

In [36], an overview of some dynamic RAM architectures is given and a performance comparison is presented. However, the results depend on the chosen CPU type and its characteristics and are therefore not easily adaptable to individual CPUs.

The most recent overview is given in [10] since dynamic RAM types such as DDR-RAMs, Direct Rambus, and SLDRAM are mentioned. This article focuses on typical applications and configurations of memory systems. However, only the performance behavior of SDRAMs is described in detail.

In [59], dynamic as well as static RAM architectures are reviewed including applications, electrical characteristics, functionality, and organization.

Thus, this report can be seen as an enhancement of [59]. It focuses on performance measures and organization details of currently available memories.

This report is organized as follows: Section 2 gives an overview of the basic functionality of memory chips which includes the description of the operation and the timing of typical RAMs. Moreover, a critical performance measure for certain applications is defined: the worst-case random access time. In section 3, the different currently available memory architectures are described. This covers various dynamic and static RAM chips as well as the corresponding memory modules. Section 4 summarizes and compares the performance, the capacity, and the organization as well as the electrical characteristics of the presented RAMs. Finally, this report ends with a conclusion and an outlook over RAM types that may be seen in the near future.

2 Functionality of RAMs

RAMs may be distinguished according to different features. They may be classified by volatileness, by organization, i.e., the number of memory banks, by the command and data interfaces, by application, by packaging, by their ability to keep electrical charge, and by other electrical characteristics. Not all of these criteria are considered in this report.

This report deals with volatile memories such as dynamic and static memories. That is, a supply voltage of typically 2.5 V to 5 V must be permanently supplied for the RAM chip to function properly. Otherwise, the contents of the memory cells are lost. Moreover, some RAMs lose electrical charge after a certain amount of time even though a constant supply voltage is provided. These RAMs are called Dynamic RAMs (DRAMs). They must refresh their charge periodically. RAMs that do not require a refresh operation are called Static RAMs (SRAMs).

Furthermore, DRAMs and SRAMs can be subdivided into synchronous and asynchronous RAMs. Synchronous ones buffer data, address, and control signals in order to separate the timing of the memory cells from the timing of the controlling processor. This way, the data and control interfaces of the memory can be configured for a pipelined operation. Asynchronous RAMs do not buffer signals. In the worst-case the RAM may stall the operation of the central processing unit (CPU) since the CPU is directly dependent on the latencies and delays of the RAM chip. Mixed architectures are also possible. For example, extended data output RAMs (EDO RAMs [59]) use a synchronous data interface but asynchronous control and address interfaces.

Since synchronous RAMs become more and more common, this report only considers this RAM type.

2.1 Organization of RAMs

A RAM consists of a multitude of memory cells which hold the information of one bit each. A certain amount of bits, typically four to 32 bit, forms the smallest accessable piece of information that can be addressed. Memory cells are arranged in a two-dimensional array. Thus, a certain piece of information can be accessed by using its row and column index (address) in the two-dimensional array. Addresses are transferred from the memory controller to the RAM and are decoded within the RAM chip in order to keep addresses short. This is done separately for rows and columns. The address pin count of a RAM chip is further reduced by transferring the row and the column address for one access successively. In this manner, a whole row within a memory array is selected first and a column is chosen thereafter.

The contents of a whole row are amplified by the so-called *sense amplifiers*. Since sense amplifiers form a considerable part of the total cost of a RAM chip, memory arrays are usually asymmetric, i.e., they have less columns than rows to keep the amount of information that must be amplified small. A memory array together with the corresponding decoders and sense amplifiers is usually called a *memory bank*.

A RAM chip may consist of several memory arrays. By using them concurrently, the capacity and the throughput of the RAM are increased. On the one hand, each memory array within a memory chip has its own row of sense amplifiers. On the other hand, the arrays in the same chip must share input and output pins and corresponding buffers.

A detailed description of the organization of RAMs with the main focus on electrical details and semiconductor structures which are needed for a DRAM operation as well as the development history of dynamic RAMs can be found in [1, 8] and an overview in [34].

2.1.1 Read accesses

If the CPU wants to read the contents of a location at a special address in the RAM, the memory controller must at first transfer the bank number and the row address of that location to the RAM. The RAM chip now decodes the row address and transfers the information of the corresponding row into the sense amplifiers. This first phase of a memory read access is often called the *activation* of a memory row.

In order to select a particular column entry in the activated row, the memory controller now has to supply the according column address to the address pins of the RAM. After decoding, the RAM is able to drive the data bus with the contents of the sense amplifiers for that column entry.

Subsequent column entries may be transferred through the data bus during consecutive clock cycles without any further control signals of the memory controller if a burst read access instruction is used. In this case, the provided column address is the starting address in the memory row for successive reads. The column address is automatically incremented by the RAM. An example for a burst read operation of eight data words is sketched in Fig. 2.

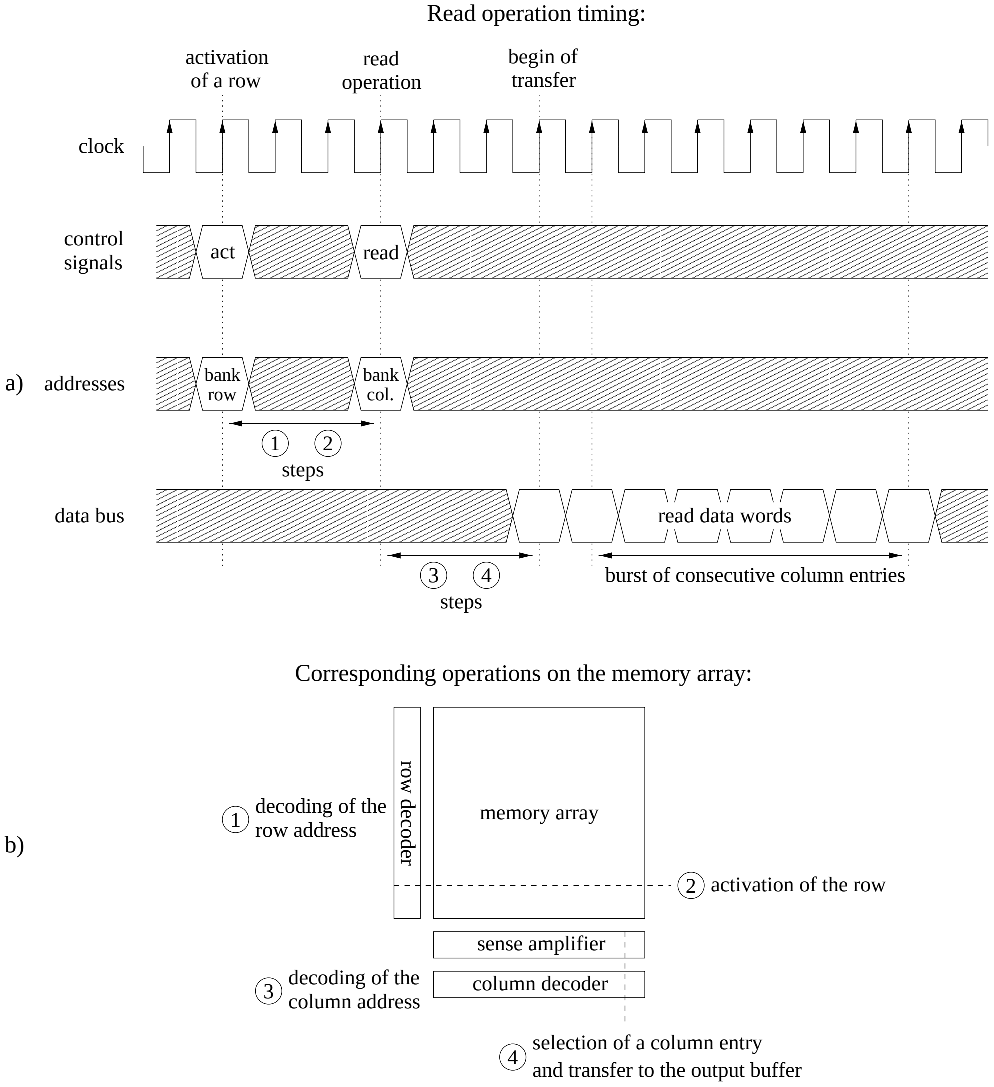

Figure 2: General functionality of a read operation: timing a), functionality b).

2.1.2 Write accesses

At the beginning of a write access, a row of the memory array must be activated in the same way as at the beginning of a read access. That is, with the help of a bank number and a row address, the information in the corresponding row is transferred into the sense amplifiers for further processing. Then, the memory controller has to transmit a column address and the data word which has to be written into a particular column entry of the activated row. This data word updates the information contents of the chosen column entry in the sense amplifiers.

As already described for the read access, a write access may also use a burst write instruction to simplify write accesses to subsequent column entries in the sense amplifiers. In this manner, column addresses are incremented automatically by the RAM and the memory controller must just drive the data pins of the RAM with different data words in consecutive clock cycles. An example for a burst write operation of eight data words is sketched in Fig. 3.

2.1.3 Precharging

Accesses to memory cells of a dynamic RAM through the sense amplifiers are destructive, i.e., the charge is lost and cannot be reconstructed by the cell itself. Hence, a precharge operation is needed to reconstruct the cell contents based on the current information that is held in the sense amplifiers. The memory controller must initiate the precharge of a memory bank before it is allowed to activate another row of the same memory array.

Some of the delay introduced by the precharge operation can be hidden if a read operation is performed before the precharge. A column entry of the sense amplifiers must be transferred into an output buffer before it can be read out by the memory controller. The precharge of the corresponding memory array may be performed concurrently to the read out of the buffer because the contents of the sense amplifiers are no longer needed.

2.1.4 Refreshing

If a dynamic RAM is used, the charge of the cells is lost slowly due to leakage effects. This is why the charge of the cells must be restored. This process is called *refresh* and implemented by an activation and a following precharge of each row of all memory arrays in the RAM. The whole device must be refreshed in intervals of typically 16 to 64 ms. Alternatively, one can refresh only a single row by time in smaller intervals of some μs in order to refresh the whole RAM within 16 to 64 ms. In this way, the penalty for refreshing the whole device is distributed over the refresh interval.

2.2 Timing

The timing requirements of synchronous RAMs depend on their architectural features. The communication mechanisms between RAMs and a memory controller can be coarsely devided into two classes: cycle-based and packet-based communication. In the former case, RAMs can accept a new instruction on the control bus with each clock cycle. Most of the currently available RAMs use this mode of communication. On the latter case, an instruction is distributed through several clock cycles in order to save some pins and due to the fact that it is usually not possible to use each clock cycle for a new instruction. In this case, instructions and bursts of data words can be seen as packets which are transferred according to a communication protocol like in communication networks.

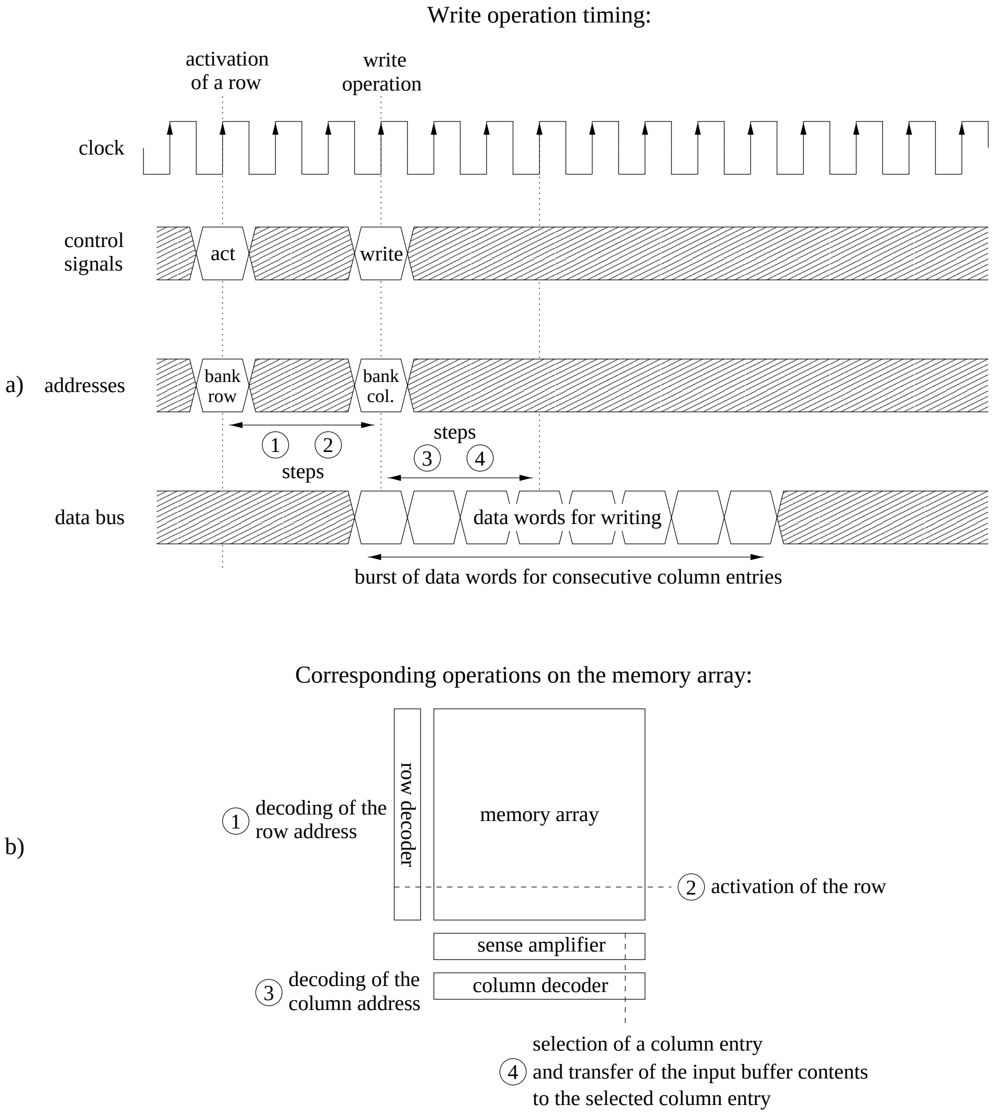
Write operation timing:
activation of a row
write operation
clock
control signals
act
write
a)
addresses
bank row
bank col.
1
2
steps
steps
3
4
data bus
data words for writing
burst of data words for consecutive column entries
Corresponding operations on the memory array:
b)
1 decoding of the row address
row decoder
memory array
2 activation of the row
sense amplifier
column decoder
3 decoding of the column address
4 selection of a column entry and transfer of the input buffer contents to the selected column entry

Figure 3: General functionality of a write operation: timing a), functionality b).

In data sheets, 30 to 40 timing parameters can be found that describe the exact behavior of the corresponding chip. In the next subsections, only around 10 are pointed out. This is sufficient to model the behavior of a RAM if it is properly used, i.e., without breaking an instruction by issuing another one too early.

2.2.1 Timing parameters for cycle-based communication

The following timing parameters are all specified in terms of clock cycles. The given values are typical for current main memory RAM chips like SDRAMs (see subsection 3.2.1). The timing parameters are additionally displayed in Fig.4.

t_{ACT}: This is the time it takes to activate a row of an idle memory bank, i.e., the memory bank is precharged and not subject to be refreshed in the next few clock cycles before the activation. After the corresponding row address has been decoded, the sense amplifiers are filled with data from the memory array. Typically, around three clock cycles are needed for this operation. This parameter is often called t_{RCD} in data sheets.

$\mathbf{t_{PRE}}$**:** This time is needed to precharge an active row. Typically, three to four cycles are spent for this instruction. The charge of the memory cells of a row is restored according to the information stored in the sense amplifiers. The memory bank is in the idle state after the precharge operation has finished. This parameter is often called t_{RP} in data sheets.

$\mathbf{t_{ROW}}$**:** This is the minimal period an activated row must be kept activated before it may be precharged, typically around 6 clock cycles. This parameter is often called the minimal RAS cycle time $t_{RAS,min}$ in data sheets and is measured from the beginning of the activate instruction to the beginning of the precharge operation for the same row. That is, t_{ROW} includes t_{ACT}.

$\mathbf{t_{AAS}}$**:** This is the minimal time between two activations of arbitrary rows in the same memory bank, typically 9 clock cycles. Thus, this parameter can also be seen as the minimal refresh interval of rows in the same bank. It consists of the activation and precharge delays t_{ACT} and t_{PRE} as well as the minimal row active time t_{ROW}. This parameter is usually called t_{RC} in data sheets.

$\mathbf{t_{AAI}}$**:** This parameter specifies the minimal time between activations of rows in different memory banks and is typically two clock cycles. That is, this parameter can be seen as the minimal time distance between two activation instructions on the control bus for interleaved activation of two rows in different memory banks. It is usually called t_{RRD} in data sheets.

$\mathbf{t_{CAS}}$**:** This parameter describes the so-called *column address strobe delay*, typically three cycles. This is the time between issuing a read instruction on the control bus and the appearance of the first data item on the output pins. In data sheets, the same parameter named t_{CAS} can be found.

burst length: The length of a burst read or a burst write operation must be specified. Typical burst lengths are four column entries or a *full page* burst. This length is the number of data word items, i.e., the number of column entries in the sense amplifier row, that are transferred consecutively without the need of providing new column addresses by the memory controller. If a full page burst is specified, the whole row of the array which can be found in the sense amplifier is read or written. RAM chips usually employ a wrap around feature which maps the access to the column after the last one in the current row to the first one of the same row. The burst length is usually configured during initialization of the RAM chip or during an idle state of the RAM, e.g., after a refresh of the whole device has been completed.

The following parameters are also displayed in Fig. 5.

$\mathbf{t_{WAR}}$**:** This is the minimal write-after-read operation delay, typically one to two cycles. That is, this time interval must be spent between the transmission of the last data word of a read operation on the data bus and the transfer of a following write instruction on the control bus. This delay is necessary because the flow direction of data words changes within the RAM. By violating this delay requirement, there is a

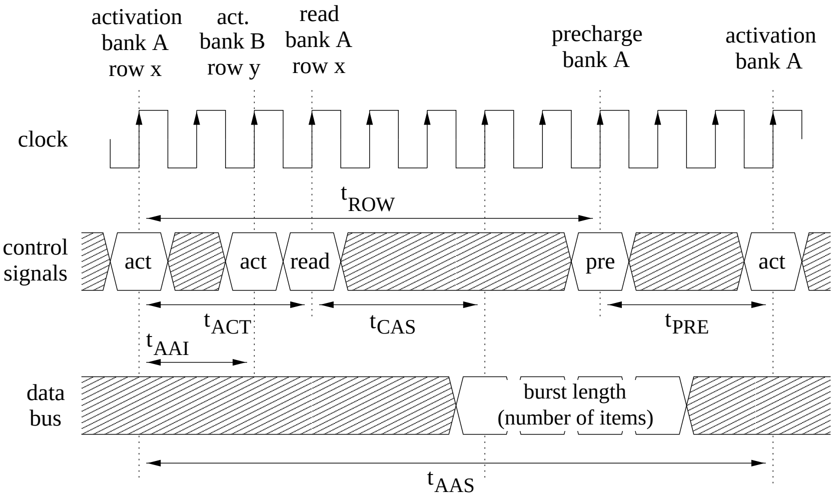

Figure 4: Timing parameters for cycle-based RAM communication.

possibility that data read out from the sense amplifiers collide with data that were already latched into the RAM chip for writing.

$\mathbf{t_{INP}}$: This delay must be spent between the transfer of the last data item of a burst write and a following precharge operation in the same bank. This delay assures that the information which is held by the sense amplifiers is up to date and that the array is precharged accordingly. This parameter is usually called t_{DPL} in data sheets.

$\mathbf{t_{WDD}}$: The supply of data words by the memory controller for writing must be delayed by this value in order to synchronize the data transfer with the decoding of the column address. RAMs, which use only a single edge per clock cycle as reference usually do not need this delay, i.e., $t_{WDD} = 0$, and the first data word for writing is supplied together with the corresponding write command. Double data rate RAMs (see subsection 3.2.5), however, usually expect a data word one clock cycle after the write instruction ($t_{WDD} = 1$ clock cycle).

$\mathbf{t_{REF}}$: The maximal refresh interval of the whole memory chip is specified by this value. The RAM chip must be refreshed completely at least once within an interval of length t_{REF}. That is, every memory row within the RAM must be activated and precharged during this interval. In data sheets, the same parameter named t_{REF} can be found.

2.2.2 Timing parameters for packet-based communication

Unfortunately, the timing parameters for SLDRAMs (subsection 3.3.1) and RDRAMs (subsection 3.3.2) are slightly different. However, the parameters they have in common are described here and differences are pointed out.

Some of the parameters are specified in the data sheets relative to the end of a particular packet. Nevertheless, the timing intervals defined in this subsection always start and end at the beginning of a packet in order to be able to compare them easily with the corresponding

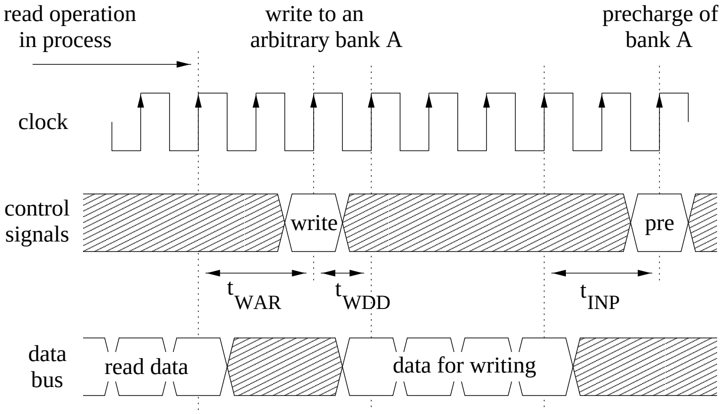

Figure 5: Timing parameters for cycle-based communication, write-after-read operation.

timing parameters of cycle-based communicating RAMs. The parameters for the packet-based case can be determined from the timing values given in the data sheets since the packet length is usually constant.

The behavior of packet-based communicating RAMs is completely different compared to cycle-based communicating ones if read-write or write-read turnarounds occur on the control bus. In the cycle-based case, the appearance of another read or write instruction on the control bus instantaneously breaks the current data transfer on the data bus, see Fig. 6 b). In the packet-based case however, new read or write instructions and data transfers may overlap without any interference (Fig. 6 a)).

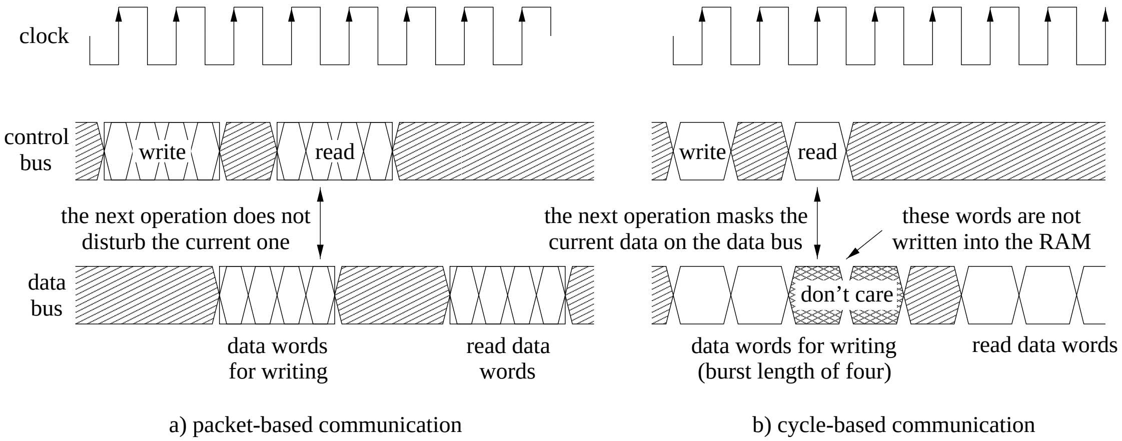

Figure 6: Effects of operation turnarounds on the control bus.

Most of the parameters have already been presented in the preceding subsection. Hence, they are only briefly mentioned here. Fig. 7 gives an overview of the introduced parameters.

$\mathbf{t_{ACT}}$**:** This is the time it takes to activate a row of an idle memory bank. This parameter is called t_{RCD} in the RDRAM data sheet and can be calculated from the $t_{BR}, t_{PR}, t_{BW}, t_{PW}$ times specified in the SLDRAM data sheet. Interestingly, the activation times are different for a read and a write operation in the SLDRAM specification.

$\mathbf{t_{PRE}}$**:** This time is needed to precharge an active row. This parameter is called t_{RP} in data sheets.

$\mathbf{t_{ROW}}$**:** This is the minimal period an activated row must be kept activated before it may be precharged. This parameter is often called the minimal RAS cycle time $t_{RAS,min}$ in data sheets and is measured from the beginning of the activate instruction to the beginning of the precharge operation in the same row. That is, t_{ROW} includes t_{ACT}.

$\mathbf{t_{AAS}}$**:** This is the minimal time between two activations of arbitrary rows in the same memory bank. Thus, this parameter can also be seen as the minimal refresh interval of rows in the same bank. It consists of the activation and precharge delays t_{ACT} and t_{PRE} as well as the minimal row active time t_{ROW}. This parameter is called t_{RC} in the RDRAM data sheet and t_{RC1} in the SLDRAM specification.

$\mathbf{t_{AAI}}$**:** (not drawn in Fig. 7) The minimal time between two activations of a row in different memory banks is specified with this parameter. It is called t_{RR} in the RDRAM case and is equal to the packet length in the SLDRAM case because an activation may be immediately followed by another activation.

$\mathbf{t_{CAS}}$**:** This delay describes the so-called *column address strobe delay.* This is the time between the issue of a read instruction packet on the control bus and the appearance of the beginning of the corresponding data packet on the data pins. In data sheets, this parameter is called t_{CAC} in the RDRAM case and t_{PR} in the SLDRAM case.

burst length: The length of a burst read or a burst write operation is specified with this parameter. The length is fixed for RDRAMs (eight items, 16 bit each) and may be set to four or eight items (16 bit) individually for each data packet in the SLDRAM case.

$\mathbf{t_{WAR}}$**:** This is the minimal write-after-read operation delay. This time interval must be spent between the transmission of a read command packet and the begin of a following write command packet on the control bus without disturbing the data bus. This parameter can be determined from the t_{CAC} and t_{CWD} times given in a RDRAM data sheet and from t_{RWD} in the SLDRAM case.

$\mathbf{t_{RAW}}$**:** (not drawn in Fig. 7) This is the minimal read-after-write operation delay. This time interval must be spent between the transmission of a write command packet and the beginning of a following read command packet on the control bus without disturbing the data bus. This parameter can be determined from the t_{WRD} parameter in the SLDRAM case. Using RDRAMs, a read command packet can immediately follow a write command packet.

$\mathbf{t_{INP}}$**:** (not drawn in Fig. 7) This delay must be spent between the end of the data packet of a burst write and a following precharge operation packet for the same bank. This parameter can be determined from the t_{WR} time in a SLDRAM data sheet and the t_{RTR} time in a RDRAM data sheet. However, the RDRAM case is described in more detail in subsection 3.3.2 since additional control packets must be used after a write command packet and the following corresponding precharge command packet.

t_{WDD}: The supply of the data packet by the memory controller for writing must be delayed by this value in order to synchronize the data transfer with the decoding of the column address. This parameter is called t_{CWD} in a RDRAM data sheet and t_{PW} in a SLDRAM data sheet.

t_{REF}: The maximal refresh interval of the whole memory chip is specified by this value. In data sheets, the same parameter named t_{REF} can be found.

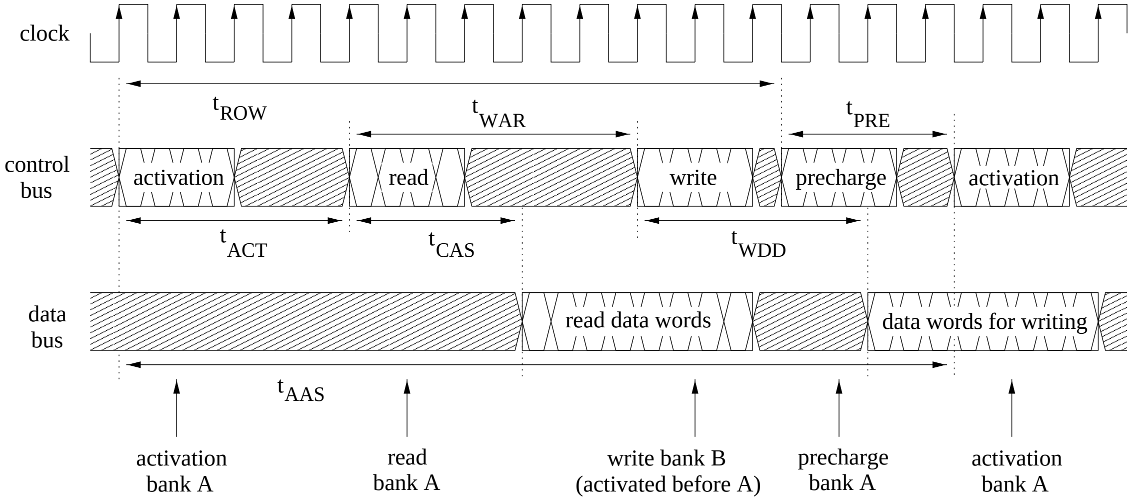

Figure 7: Timing parameters for packet-based RAM communication.

2.2.3 Worst-case access time

For random accesses, i.e., a read or write access to an arbitrary address, a RAM should deliver or store a data item as fast as possible. Unfortunately, there are situations in which a random access needs more time than in other situations since the RAM is in an unfavorable state. For instance, a particular row is activated and now another row of the same memory array is needed. Thus, the current row in the sense amplifiers must be precharged and the new one has to be activated. Furthermore, an access may be delayed due to a turnaround of the access type, e.g., a read access follows a write access, as the direction of data words on the data path in the RAM changes.

The worst-case has been identified to be a read access after a write access in different rows of the same memory array. The delay is measured from the next active clock edge after the completion time of the write access, i.e., the data bus is released but the corresponding row is still in the sense amplifiers. If the memory controller now wants to issue a read instruction for a data item of another row in the same memory array, the current row must be precharged and the next one activated. Furthermore, the delay is extended by the column address strobe delay t_{CAS}.

The approach to keep a row activated until another row is required is called *open-page policy* and works obviously best if successive operations access the same memory row. On the other hand, one may consider a *closed-page policy* which always precharges the sense amplifiers after an access. The most common memory controller chips use an open-page policy. Thus, the worst-case access time defined in this subsection considers an open-page policy, too.

Therefore, the worst-case delay of an arbitrary access is measured from the release of the data bus by the preceding write operation to the appearance of the first data item of the following read operation. This delay is given by the sum of the following latencies: the data in to precharge delay t_{INP} minus the time distance from the transfer of the last data item of the write access to the next active clock edge (typically half a clock cycle or a full clock cycle for cycle-based communicating RAMs, zero for the packet-based case), the precharge delay t_{PRE}, the row activation delay t_{ACT}, and finally the column address strobe delay t_{CAS}.

$$t_{access,worst-case} = t_{INP} + t_{PRE} + t_{ACT} + t_{CAS}$$

Up to one clock cycle must be subtracted from $t_{access,worst-case}$ if a cycle-based communicating RAM is used (see the definition for t_{INP} in that case).

One may notice, that the minimal active time of a row t_{ROW} of the preceding write operation is not taken into account, assuming that the burst length of the write instruction is sufficiently large. Thus, the active row can be precharged t_{INP} (data-in to precharge delay) after the write burst operation has been finished. This condition is usually fulfilled if a burst length of greater or equal four consecutive data items is used.

Finally, it can be seen that the worst-case access delay is determined by the delay characteristics of the memory array core and not by the speed of the memory interface, i.e., not by the delay of successive accesses to consecutive column entries of an active row in the sense amplifiers.

3 Available synchronous RAM types

Synchronous SRAMs and DRAMs have several features in common. They both have synchronous interfaces for the control bus and the input/output buses. These interfaces isolate the main memory cell arrays from the signals of the memory controller. The control interface has a command pipeline for every memory bank in the RAM. Nevertheless, it is normally not possible to transfer a memory instruction on each clock cycle for interleaved processing of the parallel memory banks because the banks in a RAM chip have to share a single data bus and an input/output buffer pair and the RAM chip itself does not check for data collisions on its internal data path.

Read and write accesses can be used in burst operation mode, that is, data words on successive addresses are transferred without the need of additional address transfers by the memory controller. A sense amplifier row is subdivided into segments of the burst length. If a burst operation does not start at a boundary of such a segment, the burst operation wraps around to the beginning of the segment when the boundary is reached in order to read or write the whole segment. So-called interleaved and linear bursts are distinguished. A linear burst counter increments the current column address modulo the burst length. An interleaved burst counter may increment or decrement the least significant bits of the column address dependent on the starting address and according to a fixed scheme. That is, only single column entries of the same row within the boundaries of a burst length segment are interleaved. Using the interleaved burst operation mode of a RAM does not mean that accesses to different memory banks are interleaved.

Furthermore, some RAMs use special control instructions to combine operations. For instance, there are control instructions which perform an activation of a row and a consecutive read or write operation. Moreover, read and write instructions can often be combined with a subsequent precharge operation.

RAMs usually offer different power-down modes to reduce the power dissipation during idle time. These modes as well as the different initialization and calibration procedures are not considered in this report.

3.1 RAMs for particular applications

3.1.1 Multibank DRAM (MDRAM)

MDRAMs [53, 68], designed by Mosys Inc. and produced by Siemens, are mainly used for graphic cards in personal computer systems. Currently, they are more and more replaced by synchronous graphics RAMs (SGRAMs, see subsection 3.2.4).

Addresses and data words are multiplexed on the same bus. The control bus uses the rising edge of the clock for identifying new instructions. Data and address transfers are performed on both edges of the clock. Unfortunately, the interface for control signals is not implemented completely synchronously since the control signals on the bus must be kept constant during the whole read or write operation, see Fig. 8 for an example. That is, read and write instructions cannot be performed in a pipelined manner. No other activation or precharge instruction can be issued during a read or write operation. Only activation and precharge operations for different banks can be performed in an interleaved way during another activation or precharge operation. Furthermore, read and write operations are always bursts and the burst length cannot be configured statically. That is, each read and write operation must be finished by an explicit stop instruction on the control bus (see Fig. 8). Thus, a clock cycle is always wasted on the control bus.

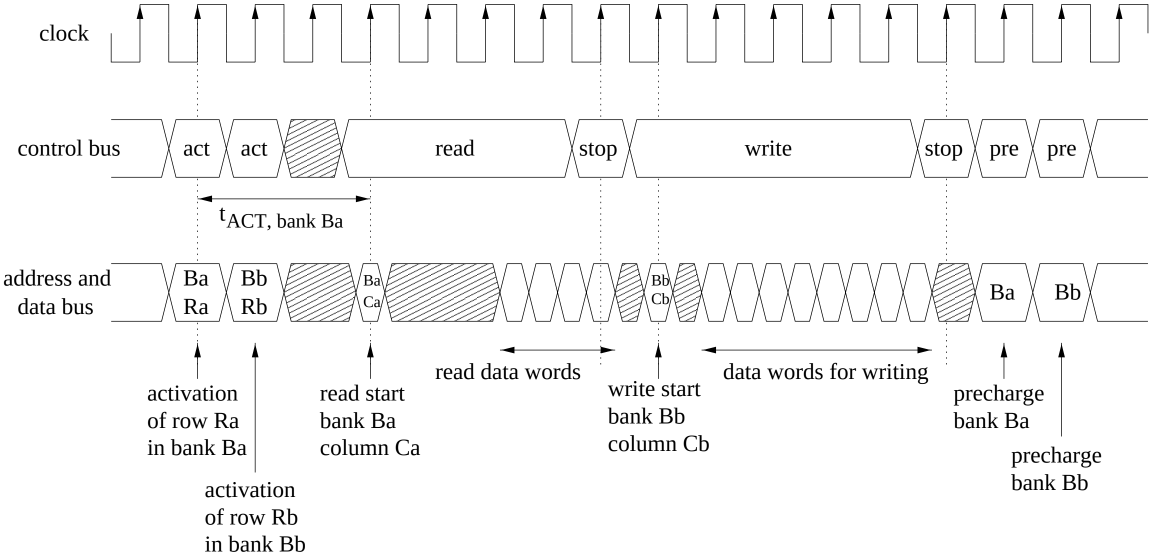

Figure 8: MDRAM timing constraints.

On the one hand, MDRAMs use a very high number of up to 80 memory banks. On the other hand, the bank size is very small (32 KByte). Thus, the chip capacity is limited to a maximal value of 2.5 MByte. Due to the small bank size and the high bank count, the row active time t_{ROW} can be kept small (four clock cycles, 24 ns) and up to 10 KByte of data can be kept available in the sense amplifiers of all banks.

Pairs of banks can be assigned to different logical address spaces. There are special registers on the chip which can be configured for this purpose.

The maximal refresh interval of 16 ms, in which the RAM must be completely refreshed, is rather long compared to other RAMs. Moreover, instructions for the concurrent activation or precharge of multiple banks are not provided and thus a long time period must be spent for refreshing all banks.

3.1.2 Cached DRAM (CDRAM)

A CDRAM [51] is a special RAM produced by Mitsubishi Corp. that consists of a single 2 MByte DRAM memory bank which is coupled with a 2 KByte SRAM through a 128 bit wide internal bus. The rising edge of the clock is used for all operations.

There are separate external data, control, and address buses. The DRAM and SRAM parts must be addressed through separate pins. That is, a CDRAM can be seen as a DRAM core with a small on-chip cache. On the one hand, the control is not transparent, i.e., the two RAM parts must be controlled as two separate devices. On the other hand, the wide internal bus reduces the communication overhead for maintaining consistent DRAM and cache contents.

There is no burst operation mode and since there is only a single DRAM memory bank, there is also no interleaved operation. The sense amplifiers of the memory bank can hold 512 Byte available.

The t_{ROW} time with seven clock cycles (49 ns) as well as the refresh cycle time of 64 ms are in the normal range.

3.1.3 3DRAM

The 3DRAM of the third generation [48] is a special RAM designed by Mitsubishi Corp. for two- and three-dimensional graphics support on graphic cards. The 3DRAM can be seen as a dual-ported RAM with different clock speeds on each port. There are separate control and address buses. Both ports and the buses use the rising edge of the clock for transfers. The first port is a read-only port and provides a constant bit stream which is intended to be used for the mandatory screen refresh of the graphic card. The second port is used for read and write accesses to the pixel values of the screen contents. The number of columns and rows of a memory bank resembles common video screen formats, namely 256 rows by 640 columns of 16 bit entries.

The 3DRAM combines a four-bank 1.25 MByte DRAM with a small on-chip ALU which can relief a two- or three-dimensional graphics rendering controller of basic calculations, such as simple blend, mask, stencil, compare, and logical functions on pixel values. Since the ALU is integrated on the chip, it is able to use a 256 bit wide internal bus for fast accesses to the sense amplifier contents of the memory banks. Furthermore, the DRAM banks and the ALU are decoupled by an additional 256 byte SRAM buffer, which is used by the ALU to load and store data values for calculations.

Since the sense amplifiers of a memory bank can only supply one access at a time, namely a transfer for the video data output port or a transfer on the internal bus, two additional small 80 Byte buffers are placed before the video data output port (see Fig. 9).

This way, video data can be constantly read out through the 16 bit wide video data port while another transfer on the internal bus is performed concurrently.

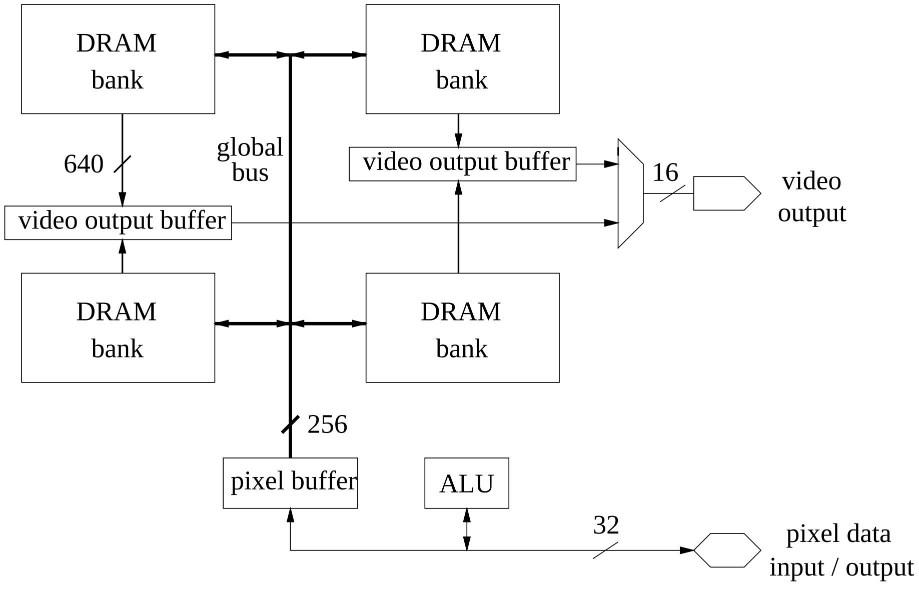

Figure 9: 3DRAM overview.

The four DRAM banks can be exploited in an interleaved fashion. Read and write operations cannot use a burst mode. However, the single accesses to the sense amplifiers use wide internal data paths, namely 256 bit to the internal bus and the ALU SRAM buffer and 640 bit to the video data port buffers.

The t_{ROW} time of eight clock cycles (80 ns) is relatively long and the 3DRAM must often be refreshed, namely every 16 ms.

3.2 Cycle-based communicating DRAMs

3.2.1 SDRAM

Currently, SDRAMs [26, 37, 65, 74, 38, 16, 22, 50, 64, 70] are the most often used synchronous DRAM type. Almost every semiconductor manufacturer produces some SDRAM variants. Especially, the RAM modules (i.e., several RAM chips mounted on a printed circuit board, see section 3.5) according to the PC 66/ PC 100 SDRAM specifications [28, 29, 27] by Intel and the imminent PC 133 standard by the Joint Electron Device Engineering Council (JEDEC) are used in personal computers as main and graphic card memory, in workstations, and in network shared memory switches.

SDRAM chips are currently most common in 64 Mbit, 128 Mbit, and 256 Mbit capacities distributed over up to four banks. That is, SDRAMs provide the largest memory banks compared to all other RAM types with a size of up to 64 MBit. Up to 4 KByte of information can be offered concurrently in the sense amplifiers of all banks. There are separate buses for data, address, and control signals, which all use the rising edge of the clock as reference. The SDRAMs of the 64 MBit, 128 MBit, and 256 MBit generations are pin compatible to each other. The additional address pins of the larger RAMs are unused in the smaller ones.

Read and write accesses can be used in a burst mode and the burst length is programmable up to a full page burst. However, this mode is not necessarily supported by all SDRAM implementations. The refresh overhead can be kept small since the maximal refresh interval of 64 ms is long compared to other RAM types and several banks can be activated and precharged concurrently with a single control instruction.

The minimal row active time t_{ROW} of five to six cycles (40 ns) is in the middle range of all compared RAM types. This statement is also true for the column address strobe delay t_{CAS} of three to four clock cycles which corresponds to 20 ns to 30 ns.

3.2.2 Enhanced SDRAM (ESDRAM)

ESDRAM [12, 21] is a slightly modified SDRAM concept by Enhanced Memory Systems Inc. which is produced by IBM Corp. The basic characteristics look very similar to SDRAM features. An ESDRAM can be used with SDRAM instructions and may therefore replace an existing pin compatible SDRAM of equal size. The ESDRAM's enhanced features are not used in this case.

The performance of ESDRAMs is increased by two design improvements. Firstly, the memory arrays have been redesigned to achieve shorter access times. Secondly, each memory bank is coupled with an additional row SRAM cache of 512 Byte. These caches can be used concurrently to the basic sense amplifiers which also keep the contents of a row of 512 Byte available. That is, two arbitrary rows of each memory array are available for fast column accesses, see Fig. 10.

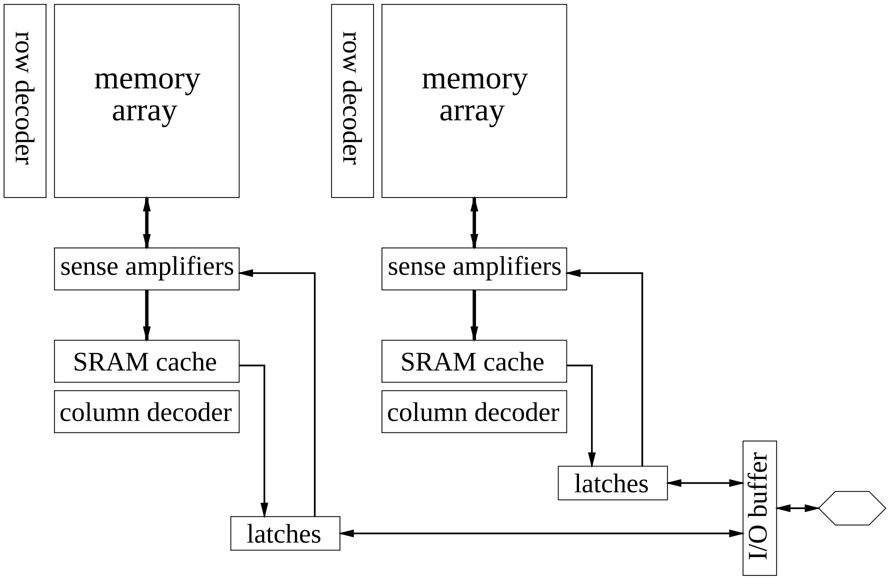

Figure 10: ESDRAM data flow path.

Unfortunately, there are limitations for the accesses. The SRAM row caches can only be used for read accesses. They are automatically updated with the sense amplifier contents when a read instruction occurs on the control bus for the row that is held in the sense amplifiers. After the update, the memory bank can be precharged and another row can be loaded into the sense amplifiers. The execution time for these operations may be completely hidden since read operations can be fulfilled by the SRAM cache contents if the content of the new row is not yet needed. Refresh operations can be hidden in the same way. However, write operations always perform on the sense amplifiers. That is, one row R1 of a memory array can be kept open in the sense amplifiers for writing and another row R2 of the same memory array is available in the SRAM cache for reading. However, in this situation column entries cannot be read from row R1 or be written to row R2 directly. The memory bank must be precharged and the corresponding row has to be reactivated for the according operation. If the row cache and the sense amplifiers hold the information of the same row, the contents of R2 are updated automatically in case a data item is written to row R1.

On the other hand, the ESDRAM size of only 2 MBytes distributed over two memory banks compares badly with current SDRAM sizes. SDRAMs are already one to two generations ahead of ESDRAMs. Nevertheless, the delay parameters are the shortest compared to all other DRAM types. A column address strobe delay t_{CAS} of only 12 ns (two clock cycles) can be achieved as well as a minimal row active time t_{ROW} of four clock cycles. Furthermore, the activation time t_{ACT} of only 12 ns reduces the refresh overhead and the worst-case random access time. Finally, the maximal refresh interval of 64 ms is the same as for ordinary SDRAMs.

3.2.3 Virtual Channel SDRAM (VC-SDRAM)

VC-SDRAM [55, 71] is another SDRAM variant which is also pin compatible with standard SDRAMs of the same size. The offered VC-SDRAMs belong to the 64 MBit generation and use two memory banks with a common row size of 512 Bytes.

The enhancement introduced with VC-SDRAMs consists of 16 small SRAM cache areas called channels. The channels have a size of a quarter of a row each, i.e., 128 Bytes. This is why each sense amplifier row of each bank is subdivided into four segments. The contents of a segment can be transferred into an arbitrary channel; see Fig. 11 for an overview. All read and write operations perform on channel contents. That is, a read operation is subdivided into two steps in the worst-case. Firstly, the contents of a segment must be copied into a channel with a *channel prefetch* operation and secondly, the channel must be read out with a *channel read* operation. The same is true for a write operation (*channel write* followed by a *channel restore* operation). The memory controller has the choice of using two distinct memory instructions for these two steps or of issuing a single SDRAM compatible instruction for the same operation. The SDRAM compatible instruction is then automatically split into the two phases by the VC-SDRAM chip controller. Two adjacent segments of the same sense amplifier row can be copied into adjacent channel numbers with a single so-called *pair prefetch* operation.

Since the memory arrays and the channels can operate concurrently, activation, precharge, and refresh penalties can be hidden completely. On the other hand, the memory controller has to take care of the data consistency of the sense amplifier and channel contents.

The burst length is programmable and a length of 16 items is provided in addition to the standard SDRAM burst lengths. A full page burst is not supported.

The delay characteristics are similar to standard SDRAMs. The device must be refreshed every 64 ms. The minimal row active time t_{ROW} of seven cycles (49 ns) is relatively long and increases the refresh overhead. The column address strobe time t_{CAS} is one clock cycle longer than the typical SDRAM t_{CAS}. This can be explained by the communication overhead introduced by the two-step read operation (channel prefetch and consecutive

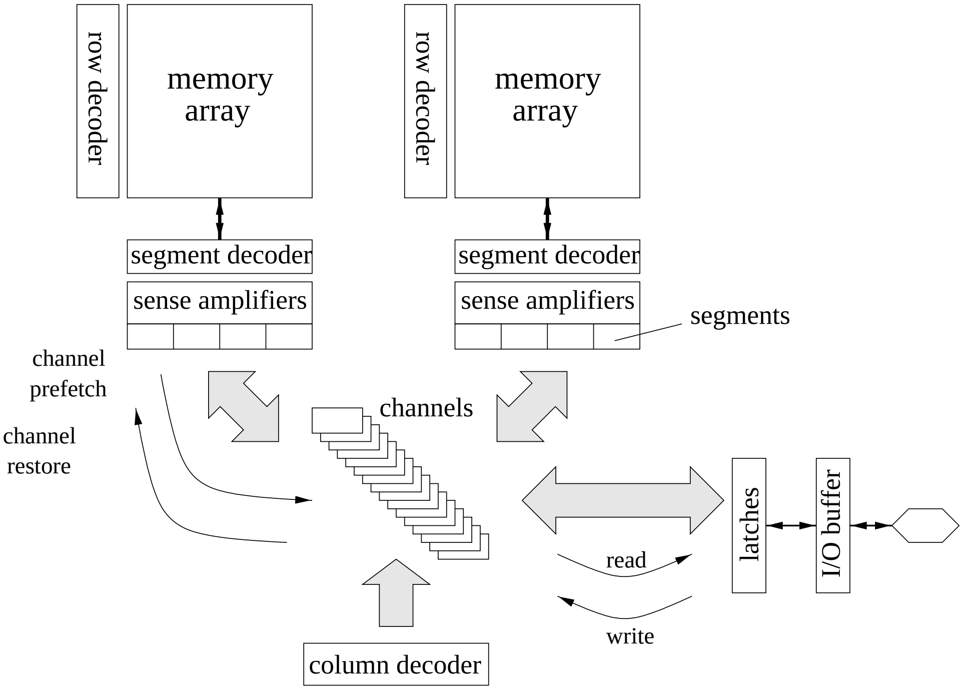

Figure 11: VC-SDRAM operation concept.

read from the channel).

3.2.4 Synchronous Graphic RAM (SGRAM)

SGRAMs [13, 17, 49, 62, 69] are produced by a variety of semiconductor manufacturers. SGRAMs can be seen as small and fast SDRAMs with additional functionality that especially supports basic graphic functions on graphic cards in personal computers. The same instruction format as for SDRAMs can be used. However, SGRAMs are not pin compatible with SDRAMs.

SGRAMs additionally have two special registers, a bit mask register and a color register, which both have the same bit width as the data bus. The bit mask register is used in the *write per bit* mode in which each single bit on the data bus can be masked during a write operation. The contents of the color register are used in the *block write* mode. In block write mode, the color register is copied eight times in eight consecutive column entries of a sense amplifier row in just one clock cycle. Additionally, the bit mask can be used.

SDRAMs and SGRAMs have their main features in common. The burst length is programmable and a full page burst is supported. There are instructions for the concurrent precharging of both memory banks. Control, data, and address buses are separate and they all use the rising edge of the clock as reference. The capacity of 2 MBytes is distributed over two memory banks. A row size of 1 KByte is common. However, the external data bus width per chip of 32 bit is larger than the bus width of ordinary SDRAMs (four to 16 bit).

With a t_{ROW} of six cycles (40 ns) and a t_{CAS} of just two clock cycles, an SGRAM is slightly faster than the average SDRAM. SDRAMs and SGRAMs must keep the same refresh interval of 64 ms.

3.2.5 Double Data Rate (DDR) SDRAMs and SGRAMs

DDR-SDRAMs [23, 43, 52] and DDR-SGRAMs [44, 61] are high throughput variants of SDRAMs and SGRAMs. They use the rising edge for instruction transfers on the data bus. Contrary to SDRAMs and SGRAMs, DDR-RAMs use both the rising and the falling edge for data transfers in order to double the throughput of the input-/ output interface. Non-DDR RAMs are often called Single Data Rate (SDR) RAMs to distinguish them from their DDR variants.

An additional signal called *data strobe* must be used in order to synchronize the data transfers with the edges of this signal. Thus, it can be seen as an additional clock signal for the data flow which must be driven by the memory controller for write operations and by the memory chip for read operations. An example of an arbitrary read operation followed by an arbitrary write operation in different rows of the same memory bank is given in Fig. 12. Furthermore, the behaviors of an SDR and a DDR RAM, which are controlled by the same signal sequence on the control bus, are compared for this particular access sequence.

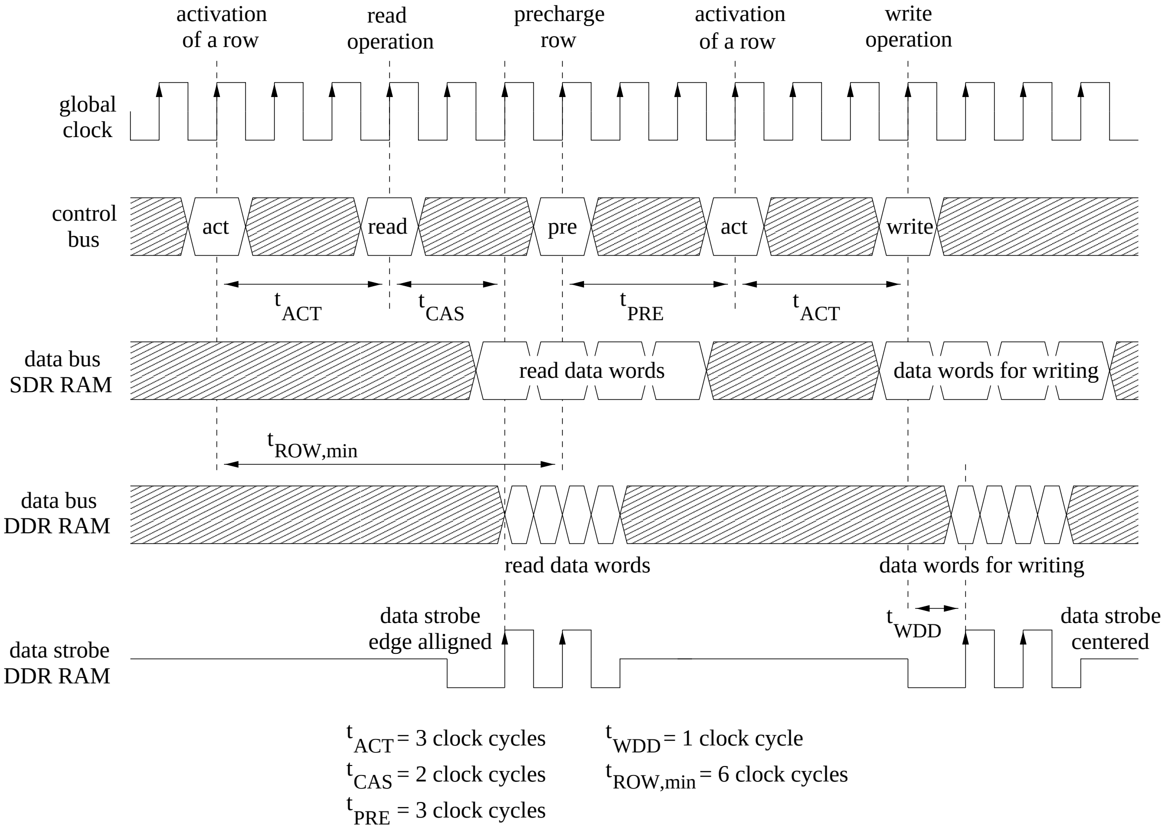

Figure 12: SDR and DDR access mode comparison.

DDR-SGRAMs and SGRAMs characteristics look very similar. They both use programmable burst lengths, the same instruction format, a 32 bit wide data bus, and the same package with almost the same pinout. Differential clock inputs are used and some address pins are interpreted differently, since a DDR-SGRAM uses twice the number of banks which are on the other hand of half the row size.

The basic timing parameters of DDR-SGRAMs and SGRAMs such as t_{ROW} and t_{CAS} are about the same. Therefore, the minimal row active time can be a throughput limiting factor for DDR-SGRAMs since the data transfer time is halfed, but the rows must be kept activated for the same amount of time than in SGRAMs. See again Fig. 12 for a situation in which a read and a write operation cannot be performed in an interleaved fashion. In this case, the DDR variant is not at all faster than the SDR version due to the minimal row active time t_{ROW}.

Moreover, the maximal refresh interval of DDR-SGRAMs is reduced by a factor of up to four compared to SDR-SGRAMs of the same size and clock frequency. This causes a bigger refresh delay penalty.

The comparison of DDR-SDRAMs and SDRAMs basically underpins the same statements. The main difference between the two types of memories is the speed of the input-/ output interface. The characteristics of the underlying memory core are the same. The core is not faster than in the non-DDR generation. Therefore, the minimal row active time becomes a potential performance bottleneck.

Furthermore, the packages of DDR-SDRAM and SDRAM memory types are different since for a DDR-SDRAM more supply pins are needed and the clock must be provided through differential inputs.

A detailed description and comparison of DDR SDRAMs and SGRAMs can be found in [2].

3.3 Packet-based communicating DRAMs

3.3.1 Synchronous-Link DRAM (SLDRAM)

SLDRAM [14],[72, 40] is a memory specification according to future IEEE (P1596.7) and JEDEC (VSMP, TSOP packaging) open standards. It includes the description of an I/O bus for interconnecting up to eight SLDRAM devices with a single memory controller. The I/O bus scheme needs two distinct clocks for data transfers and a third clock signal for control signals. All clocks use differential input pins. The data clocks can be driven by the controller for write transfers to the RAM. The data clocks may be driven by the SLDRAM devices for read transfers to the controller. This is why SLDRAMs need a complex calibration procedure for read, write, and clock delays at the startup of the whole memory system.

Control and address signals share a single bus; the data bus is separate. Both buses use both edges of the clock for transfers. Read, write, activation, and precharge request packets are transferred on four consecutive clock edges. A burst data transfer may consist of four or eight data words which are transmitted in four and eight consecutive clock edges, respectively. The according burst length can be chosen separately for each request packet. In order to reduce the usage of the control bus, an activation, a read or a write instruction, and a following precharge operation can be specified in a single request packet.

A 64 Mbit SLDRAM consists of eight banks with a relatively large row size of 1 KBytes. That is, a total of 8 KBytes of data can be kept available in the sense amplifiers of all banks.

The maximal refresh interval of 64 ms is common for a synchronous DRAM. On the other hand, the minimal row active time t_{ROW} of 60 ns corresponding to 12 clock cycles is not decreased compared to other DRAMs. Since SLDRAMs use both edges of the clock and a data packet maximally needs four clock cycles, the t_{ROW} time may be a performance limiting factor for random accesses. A minimal t_{CAS} time of 36 ns (7.5 clock cycles) and a t_{ACT} time of 30 ns (six clock cycles) resemble the usual delays of a standard DRAM core. Interestingly, the minimal activation delay seems to be smaller for activation instructions followed by a write than for an activation instruction followed by a read.

3.3.2 Direct Rambus DRAM (RDRAM)

RDRAM [5], [60, 18, 19, 20, 73] is a memory specification developed by Rambus Inc. Royalties must be paid if a semiconductor manufacturer wants to use the RDRAM architecture. However, since some of the biggest personal computer and consumer electronics suppliers and manufacturers such as Intel, AMD, Sony, Acer, and Compaq have decided to use RDRAMs in their future products, RDRAMs will certainly play a major role in the DRAM business.

RDRAMs will be initially available in 64 Mbit, 128 Mbit, and 256 Mbit sizes. The 64 Mbit variant uses 16 memory banks, the bigger sizes 32 banks. The row of the memory banks have a size of 1 KByte. However, the sense amplifier rows cover just half of the row entries carrying only 512 Byte of information. Therefore, sense amplifier rows must be shared between adjacent memory banks in order to cache the information of a complete memory array row in two sense amplifier rows. This is why just half of the available memory arrays can be kept activated concurrently in the best case with the additional constraint that adjacent memory banks of an activated bank must remain in a precharged state and cannot be activated. On the other hand, the sense amplifiers are still able to cache 8 KByte of information in a 64 Mbit chip. This is a relatively large value compared with the other DRAM types.

A RDRAM uses a single bus for addresses and control instructions as well as a separate bus for data transfers. Both edges of the clock are used on both buses. There are two differential clock signals for each direction of the signal flow. The clock signals must be provided externally. The transmission of all kinds of packets is started at the falling edge of the clock and needs four clock cycles (eight clock edges) for completion. Thus, the burst length is fixed to eight data words of 16 bit. The maximal clock frequency of 400 MHz is the highest one of all compared RAM types.

Additional instructions are necessary for the control of the write buffer. This buffer is always used if data is written to the RDRAM. The buffer carries the information of a single burst write transfer. The memory controller must take care of the data consistency of the write buffer and the sense amplifier contents. That is, the contents of the buffer must be transferred to the corresponding sense amplifiers with the help of a so-called *retire* instruction packet after a minimal delay t_{RTR} after the beginning of the corresponding write command packet and before the appearance of any subsequent read, write, or precharge instructions for the same memory location.

However, a retire instruction can be skipped if a read or a write instruction to another RDRAM chip is issued after the time t_{RTR}. In this case, the write buffers of all other devices are retired automatically. Moreover, a write instruction t_{RTR} after another write to the same device also retires the write buffers automatically.

Nevertheless, the additional retire commands may delay subsequent operations. Consider a situation where a read operation follows a write operation to the same memory address. In this case, the read operation must be delayed until the write buffers are in a retired state.

The maximal refresh interval of 32 ms is in the middle range of the compared memory types. The minimal row active time t_{ROW} of 50 ns (20 clock cycles) as well as a column address strobe delay t_{CAS} of 30 ns (12 clock cycles) are typical for a DRAM core of the 64 Mbit generation. Therefore, the t_{ROW} time is a limiting factor for fast arbitrary accesses through the "overclocked" RDRAM input-/output interface. Furthermore, there are no instructions for the concurrent activation or precharge of parallel memory banks within a RDRAM. Thus, the overhead for the refresh of the whole chip is increased.

3.4 SRAMs

3.4.1 Pipelined burst SRAM (PBSRAM)

PBSRAMs [24, 47, 42, 39] use the same operation modes as synchronous DRAMs. Read and write instructions may be used in burst mode, the synchronous interface is pipelined and thus allows to issue memory instructions with every clock cycle. The most common sizes are currently 128 KByte, 512 KByte, and 1 MByte per SRAM chip. The main application is in external second-level cache memories of computer systems.

The buses for data, addresses, and control signals are separate. They all use the rising edge of the clock as reference. The burst length is fixed to four data items.

Since PBSRAMs are optimized for speed, the address of a data item in memory is not split into row and column addresses but fully specified in one clock cycle. The pin count is increased but an arbitrary data item can be read out with a delay of just two clock cycles independently of the address and the state of the SRAM. An equivalent DRAM ought to have the timing parameters $t_{ACT} = 0$, $t_{CAS} = 2$ cycles, $t_{ROW} = 0$, and $t_{PRE} = 0$.

Due to SRAM technology, a refresh operation is not necessary and there is no limiting minimal row active time parameter that must be kept.

On the other hand, the worst-case power dissipation of an SRAM in operation is higher than for any DRAM type.

3.4.2 PBSRAM variants

Since a PBSRAM needs two clock cycles to read out the contents of an arbitrary address after the read instruction is issued and write data must be supplied instantaneously with the write instruction, NOP instructions must be inserted if the sequence of instructions changes from read to write accesses. Furthermore, the data bus cannot be completely exploited if a read access follows a write access.

Therefore, a gain in throughput can be achieved by delaying the supply of the write data relatively to the write instruction on the control bus by the same amount of time as the SRAM delays the read out of data relatively to the read instruction. See Fig. 13 for an example of an alternating read-write access sequence. This mode of operation is called *late write*. In the PBSRAM case, cycles on the data bus are unused if write data words are not delayed. The *Zero Bus Turnaround* (ZBT) SRAM [45] by Micron as well as the *no-turnaround* SRAM [63] by Samsung use the late write mode for reducing the number of unused data bus cycles.

3.4.3 DDR-PBSRAM

DDR-PBSRAM [25, 41] are the DDR enhancement of PBSRAMs, i.e., the address and data buses use both edges of the clock. DDR-PBSRAMs are mainly used as second level cache memories.

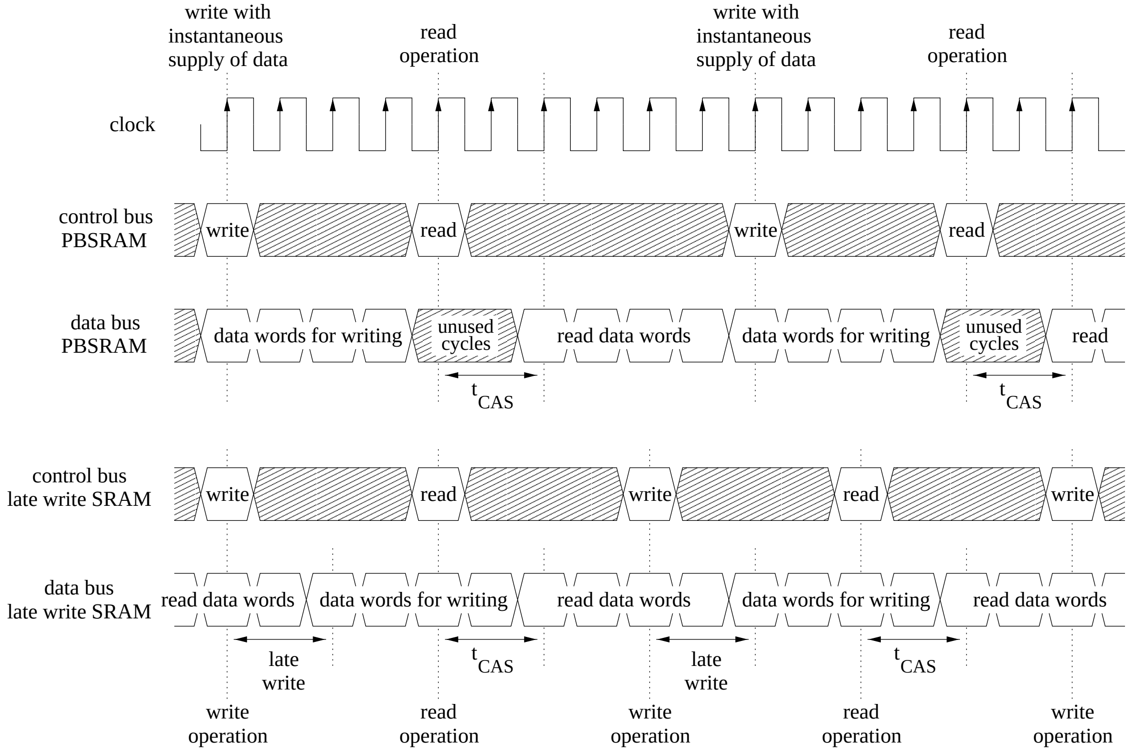

Figure 13: Influence of the late write operation mode, burst operations.

The package differs from PBSRAMs of the same size, since additional supply voltage pins are needed as well as differential clock inputs.

Since DRAM timing parameters like t_{ROW} are meaningless in the SRAM case, the throughput is improved due to halfed data transfer times for arbitrary read or write accesses. Besides, the read access delay is further reduced to 1.5 clock cycles.

3.5 Available RAM modules

A module is a printed circuit board (PCB) on which several memory chips are mounted. In addition, there is a non-volatile device on the module which can be read out by the memory controller and which determines the timing and electrical characteristics of the used memory chips. There are several module specifications available which provide a fixed data bus width. For instance, a 64 bit wide module for SDRAMs may consist of eight RAM chips with an eight bit organisation or of four chips with an 16 bit organisation. There are limitations which consider the maximal count of memory modules that can be supplied by a single memory controller due to load and timing constraints.

The most common module specifications are:

168-pin DIMM: (Dual In-line Memory Module) This module standard is specified by JEDEC [30]. It is used for a 64 bit wide data path and usually equiped with SDRAMs. If 4 bit organized RAM chips are used, the module has additional buffers in order to reduce the load of select and clock signals. Several chips are needed in parallel to drive the data bus at a time. SLDRAMs will also use this module type.

184-pin DIMM: This 64-bit wide module is used for the double data rate variants of SDRAM and SGRAM chips. The additional pins are needed for further supply pins and data strobe clock signals.

144-pin SODIMM: (Small Outline DIMM) This is a module standard by JEDEC intended for small scale RAM chip packages with a low power dissipation. These modules are mainly used in notebooks and small scale embedded systems and have a data bus width of 64 bit.

184-pin RIMM: This module type specified by Rambus Inc. will be used to combine up to 16 RDRAM chips on a single PCB. A RIMM has the same form factor as the corresponding DIMM, but is not pin compatible. The module has a 16 bit wide data bus. Since RDRAM chips also use a 16 bit data path, a single chip drives the whole bus at a time.

144-pin SORIMM: Rambus equivalent of the 144-pin SODIMM.

The modules are used for different kinds of bus topologies. DIMMs are connected in parallel to the memory controller. In order to cope with the load, some signal lines are doubled. On the other hand, RIMMs are connected serially and form a so-called Rambus *channel*. This is why unused RIMM connectors must be populated with a so-called continuity module which just connects the input data bus segment with the output bus segment of the RIMM connector.

Clock signals are used in different ways. DDR-SDRAMs and DDR-SGRAMs as well as SLDRAMs drive data strobe signals in case of a read operation relatively to a global clock signal. On the other hand, the data strobe is driven by the memory controller in case of a write operation. Thus, for each direction of data flow, a distinct clock signal is used. RDRAMs however use two clock signals that must be provided by the interface. No data strobe signal is generated by a RDRAM itself. There is a separate clock generator which drives a clock signal in one direction on the Rambus channel. At the end of the bus, the clock line is folded with another clock line which drives a clock signal in the opposite direction. That is, both data strobe signals are driven by the same clock chip.

On the serial Rambus channel, signal changes can appear in less than 2 ns. Hence, signals may have a considerable round-trip time. In [46], this time was determined to be around 10 ns in the worst-case. This has a strong influence on the read latency and the delay introduced by read-write or write-read operation turnarounds. For example, a fully loaded SLDRAM system just uses one half of the Rambus clock frequency and one third of the maximal bus length of a Rambus channel. Therefore, the worst-case round trip time of a signal is reduced and does not have such a strong impact on the performance than in a Rambus system ([46]).

The capacity limit of a Rambus channel is reached if 32 RDRAMs distributed over three RIMMs are connected to the channel. However, the storage capacity of a single channel may be further increased by using buffered RIMMs. Thus, the load per pin can be reduced but an additional delay of at least one clock cycle is introduced. Nevertheless, one may combine several separate channels to increase the capacity further. Due to the organisation of a Rambus memory system in a narrow channel with serially interconnected RDRAMs, a lot of memory banks can be used concurrently. Each RDRAM consists of at least 16 memory banks. Since just a single RDRAM drives the data bus at a time, up to 512 memory banks may be controlled individually in a fully configured Rambus system in order to hide activation and precharge latencies.

Three to four DIMMs can typically be found in a fully configured memory system consisting of up to 32 SDRAMs. Server memory controllers support up to eight DIMMs. On each DIMM, at least four memory chips have to drive the data bus in parallel. Thus, these four chips are controlled in the same way. Their memory banks cannot be controlled individually. Therefore, having four banks per SDRAM, a total of just 16 individual memory banks may be seen in a SDRAM memory system. The storage capacity may be further increased by using buffered modules or by using several DIMM data paths in parallel.

The idea to interconnect RAM chips serially to form a Rambus channel like system could also be implemented with common memory chips such as SDRAMs. However, error correcting code algorithms used in server computers take advantage of the parallel structure of DIMMs. Currently, no algorithms and storage patterns are known which perform error correction for a Rambus channel on a bit and on a device level as efficient as current algorithms on ECC-DIMM populated server systems do.

4 Comparison

A lot of characteristics can be defined by which memory chips are distinguishable. At the end of this section an overview of the characteristics of the compared RAM types is given. Table 1 summarizes the most important performance measures, table 2 shows the usual memory capacities, and finally table 3 presents the electrical characteristics as well as the packages of the compared RAM chips. In the following subsections, some of the features are explained in more detail.

4.1 Throughput of the memory interface

This value is determined by the clock frequency of the different buses for control, addresses, and data words. Most of the compared RAMs use the same frequency for all types of buses. The DDR-variants of SRAMs, SGRAMs, and SDRAMs double the maximal data throughput by using both edges of the clock for the data bus. Finally, packet-based communicating RAMs transfer all kinds of signals on both edges of the clock. Clock frequencies in the range of 66 MHz up to 400 MHz are currently provided with a main focus between 100 MHz and 200 MHz, see Fig. 14.

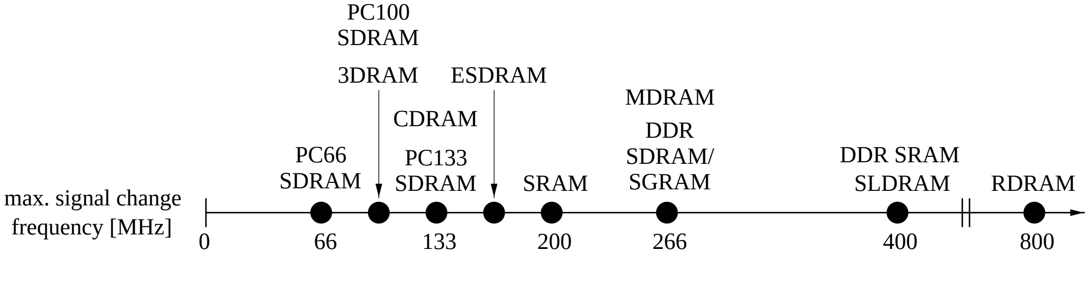

Figure 14: Maximal frequency of signal changes on all buses.

The throughput of the RAM interface can be seen as another performance measure. Long burst transfers, e.g., cache line fills, are especially determined by this measure since consecutive column entries cached in the sense amplifiers can be read out or written with the interface frequency. This frequency does not directly depend on the timing behavior of the memory core. Delays introduced due to activation and precharge latencies are neglectable if the burst length is chosen long enough.

Data bus widths of 4, 8, 16, or 32 bit are most common for a single RAM chip. Usually, a chip of a fixed capacity is offered with different data bus withs. That is, the more bits are spent for the width of the data bus, the less column entries are available in an activated row since the row size is fixed. Therefore, different configurations are possible to achieve a particular memory system of a certain size. For instance, using two 16 bit chips of the 64 Mbit generation result in a capacity of 16 MByte for a 32 bit wide system, whereas four 8 bit chips would result in a 32 bit wide system as well, but with a size of 32 Mbyte. The throughput of both systems would be the same.

Most of the RAMs are available at least with a data width of 16 bit. A width of 32 bit is most common for SRAMs and SGRAMs. The smaller data widths of 4 and 8 bit are mainly used for SDRAMs, since SDRAMs are primarily utilized for large main memory capacities.

SRAMs, RDRAMs, and SLDRAMs are also offered with data widths of 18 instead of 16 bit and 36 instead of 32 bit, respectively. The additional bit lines may be used to implement error correcting codes. For that purpose, one additional bit is spent for every data Byte.

4.2 Worst-case random access delay

This value is calculated according to the description in section 2.2.3 and mostly depends on the timing characteristics of the memory array core. Control, data, and address buses clocked at high rates do not provide a faster access. All access delays of DRAMs are of the same magnitude. Only ESDRAMs are slightly faster than the usual DRAMs, see Fig. 15.

Nevertheless, access times of SRAMs are one magnitude smaller than DRAM access delays.

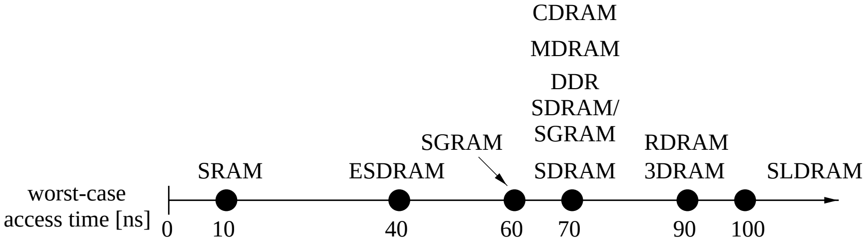

Figure 15: Comparison of the worst-case random access delay.

4.3 Capacity per chip and memory bank

A variety of chip capacities and memory bank sizes is available, see Fig. 16 for an overview. RAM chips intended for graphic applications like 3DRAMs, SGRAMs, and MDRAMs use relatively small memory banks in order to reduce activation delays and to have several banks available for latency hiding. RAMs for main memories such as SDRAMs and RDRAMs use the largest memory bank sizes. However, RDRAMs use four times more memory banks than a corresponding SDRAM of the same size. That is, more banks can be controlled concurrently with the potential to hide most of the activation and precharge delays. This is mandatory since the RDRAM memory interface is noticeably faster than the SDRAM interface and thus the sense amplifier contents are read out more rapidly. Finally, SRAMs are realised with a single memory bank.

Some RAM types, e.g. SRAMs, RDRAMs, and SLDRAMs, are available in slightly larger capacities as the ones shown in Fig. 16, because an additional bit is spent per storage Byte in order to implement error correcting codes. This 12.5% increase in capacity may also be used for normal storage of data.

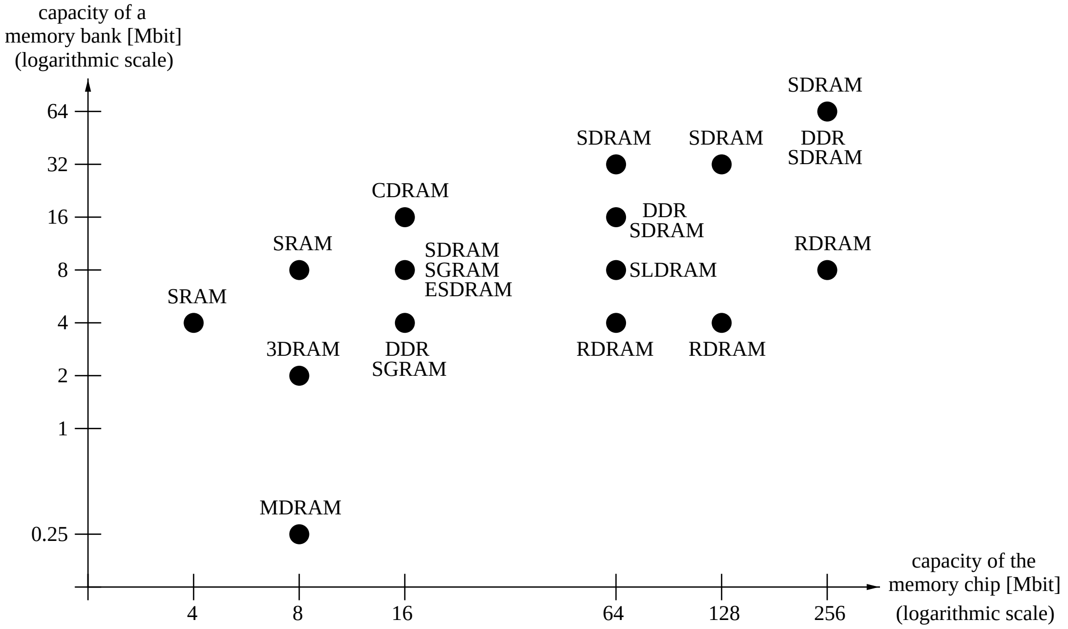

Figure 16: RAM and memory bank capacities.

4.4 Price per MByte

Since customer prices for semiconductor devices are continuously changing, this subsection just gives an impression of the relative proportions of the RAM prices to each other. The prices were specified by resellers and distributers for small quantities of chips at the end of 1998. An overview of the prices is given in Fig. 17.

RDRAMs are expected to have a 40 % to 50 % price penalty compared to a PC 100 SDRAM of the same capacity when they will be introduced in the second half of 1999. That is, the price per MByte for a RDRAM is an estimation rather than a specification by a company or distributor.

4.5 Electrical characteristics and package

The electrical characteristics such as power dissipation, supply voltage, and interface specifications as well as the package outlines are not pointed out in whole detail in this report. However, since these specifications may be needed for the choice of a suitable RAM chip, they are listed in table 3.

The complete specifications of the chip packages such as TSOP and TQFP with pin counts, spacing, etc. can be found in [33, 30]. The features of the interface types such as

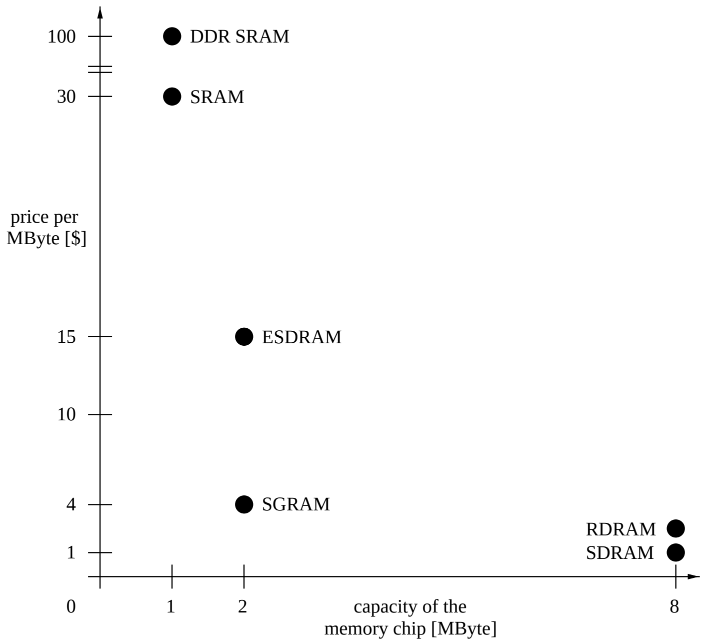

Figure 17: Price per MByte and RAM.

LVTTL and SSTL are described in [32, 31]. The power dissipation values were taken from the data sheets.

RAM chips of the same type and capacity are sometimes available with different interface types. For instance, there are SRAMs according to the LVTTL or the SSTL specification. Hybrid solutions are also possible. Certain DDR-SDRAMs for example may use LVTTL voltage levels for the control and address bus but may transfer data according to the SSTL 2 specification.

4.6 Comparison summary

Table 1 summarizes the performance features of all compared RAM types. The interface frequency is the frequency of the main clock. In the next column, the clock edges are specified which are used as reference for control and data signals. If different settings are used for control and data signals, an entry can be found for each case in the table. The refresh penalty is deduced from the minimal delay between two adjacent row activations on the same bank t_{AAS} and the number of rows and banks within a chip. Control signals for the concurrent activation and precharge of rows on different banks are exploited (if available). The worst-case (WC) latency is given by the worst-case random access time defined in subsection 2.2.3. The column *burst modes* states the number of adjacent column entries that can be transferred consecutively without the support of additional addresses from the memory controller. A *full page* burst reads out the contents of the whole activated row.

Table 2 displays the number and size of the rows, columns, and memory banks within a chip. Furthermore, the resulting total capacity is shown as well as the width of the external data bus. It can be seen that there are RAM types such as RDRAMs and SLDRAMs which are only able to address coarse column entries. The addressed column entry is then

memory type	interface frequency [MHz]	edges used [r(ising)/ f(alling)]	refresh penalty [% of cycles]	WC latency [cycles/ns]	burst modes [#items]
MDRAM	166	ctrl: r data: r+f	< 6.2 %	11 / 66	full page
CDRAM	143	r	< 0.7 %	10 / 70	none
3DRAM	83 (video) 100 (pixel)	r	< 0.8 %	9 / 90	full page (video) none (pixel)
SDRAM	143	r	< 0.6 %	10 / 70	1,2,4,8 (full page)
ESDRAM	166	r	< 0.2 %	7 / 42	1,2,4,8,full page
VC-SDRAM	143	r	< 2.4 %	9 / 63	1,2,4,8,16
SGRAM	143	r	< 0.4 %	9 / 63	1,2,4,8,full page
DDR SDRAM	143	ctrl: r data: r+f	< 1.0 %	9 / 72	2,4,8
DDR SGRAM	166	ctrl: r data: r+f	< 0.9 %	10 / 70	2,4,8,full page
SLDRAM	200	ctrl: r+f data: r+f	< 1.2 %	21 / 105	4,8
RDRAM	400	ctrl: r+f data: r+f	< 1.8 %	34 / 85	8
PBSRAM	200	r	-	2 / 10	4
ZBT/Nt-SRAM	150	r	-	2 / 13.5	4
DDR SRAM	200	ctrl: r data: r+f	-	1.5 / 7.5	4

Table 1: Performance measures of current memories.

transferred with the help of a burst access. On the other hand, e.g. SDRAMs can address fine column entries so that several consecutive column entries may be transferred within a burst access. That is, RDRAMs use addresses which are multiples of the burst transfer length.

Table 3 summarizes the electrical characteristics as well as the packages of the compared RAM chips. The detailed packages descriptions can be found in [33, 30]. The input/output interface (I/O IF) specifications are pointed out in [32, 31].

Interestingly, an SRAM may consume more power than a DRAM if it is used intensively. Moreover, the faster the device is, the lower is the power supply voltage it must cope with in order to maintain sharp signal edges.

5 Conclusion

The fastest DRAM cores are produced by Enhanced Memory Systems (ESDRAM). In addition, the option to have two memory pages of the same memory bank open at the same time within an ESDRAM may be interesting for certain applications because page turnaround times within a memory bank can be dramatically reduced.

RDRAM and SLDRAM as well as all DDR solutions are high speed interfaces wrapped around a standard DRAM core, i.e., they are not able to access the DRAM core any faster. Thus, the minimal row active time of a DRAM core becomes more and more a critical, performance limiting factor for random accesses. On the one hand, the high speed

	organization per chip			
memory type	#rows/ row size [Byte]	#columns/ column size [bit]	#banks/ total size [MByte]	bus width per chip [bit]
MDRAM	256 / 128	32 / 32	36 / 1.125	16
CDRAM	4096 / 512	32 / 128	1 / 2	16
3DRAM	256 / 1280	16 / 640	4 / 1.25	16 (video) 32 (pixel)
SDRAM:				
64 Mbit	4096 / 1024		2 / 8	
128 Mbit	8192 / 1024	2048 - 512 / 4,8,16	2 / 16	4,8,16
256 Mbit	8192 / 1024		4 / 32	
ESDRAM	2048 / 512	1024 - 256 / 4,8,16	2 / 2	4,8,16
VC-SDRAM	8192 / 512	256 - 64 / 4,8,16	2 / 8	4,8,16
SGRAM	1024 / 1024	256 / 32	2 / 2	32
DDR-SDRAM:				
64 Mbit	4096 / 512	1024 - 256 / 4,8,16	4 / 8	4,8,16
256 Mbit	8192 / 1024	2048 - 512 / 4,8,16	4 / 32	4,8,16
DDR-SGRAM	512 / 1024	256 / 32	4 / 2	32
SLDRAM	1024 / 1024	128 / 64	8 / 8	16
RDRAM:				
64 Mbit	512 / 1024		16 / 8	
128 Mbit	512 / 1024	64 / 128	32 / 16	16
256 Mbit	1024 / 1024		32 / 32	
SRAM:				
4 Mbit	256k x 16 bit		1 / 0.5	16
	128k x 32 bit		1 / 0.5	32
8 Mbit	512k x 16 bit		1 / 1	16
	256k x 32 bit		1 / 1	32

Table 2: Capacities of current memories.

memory type	chip package-#pins	I/O IF type	supply voltage	WC power consumption
MDRAM	PQFP-128, PLCC-68	(LV)CMOS,SSTL	5 or 3.3 V	0.8 W
CDRAM	TSOP-II-70	LVTTL	3.3 V	1.5 W
3DRAM	QFP-128	LVTTL	3.3 V	1.6 W
SDRAM	TSOP-II-54	LVTTL	3.3 V	0.7 W
ESDRAM	TSOP-II-44/54 (4,8/16 bit)	LVTTL	3.3 V	0.5 W
VC-SDRAM	TSOP-II-54	LVTTL, SSTL 2	3.3 V	1.0 W
SGRAM	(T/P)QFP-100	LVTTL, SSTL	3.3 V	1.1 W
DDR-SDRAM	TSOP-II-66	LVTTL. SSTL 2	3.3 V	1.0 W
DDR-SGRAM	TQFP-100	SSTL 2	3.3 V	1.6 W
SLDRAM	VSMP/HSMP-64,TSOP-II-80	LVCMOS, SSTL 2	2.5 V	1.3 W
RDRAM	μBGA-74/126	Rambus-SL, 1.8 V	2.5 V	TBD
SRAM	TQFP-100,Bump BGA-119	LVTTL, SSTL 2	3.3 V	1.6 W
DDR-SRAM	Bump BGA-153	HSTL, LVTTL	2.5 or 3.3 V	2.5 W

Table 3: Electrical characteristics and packages of current memories.

interfaces may help to save I/O pins as the time to transfer a certain amount of data from the interface is reduced. On the other hand, maintaining signal integrity is much more difficult at 400 MHz and needs additional power, ground, and clock pins. Finally, RDRAM technology must be licensed. Nevertheless, since most of the personal computer suppliers and manufacturers have decided to support RDRAMs in the future, Rambus RAMs are likely to dominate the PC market in the near future.

SDRAMs and SGRAMs, which are actually very similar products, are currently widely used. Therefore, their corresponding module standards DIMM and SODIMM are common and SG/DRAM populated DIMMs may be purchased in a variety of configurations.

SRAM technology is the fastest RAM access solution at the expense of a cost and power dissipation penalty. However, designing a memory controller for synchronous SRAMs is easier than designing one for a synchronous DRAM since all the difficulties introduced by using DRAM technology such as refreshing intervals and access delays dependent on the internal state of the RAM do not appear.

6 Outlook

This section tries to give a short overview of the development of RAM chips that may be seen in the near future.

First of all, double data rate variants of the ESDRAM and VC-SDRAM types are already manufactured. Moreover, prototypes of the next generation DDR-SDRAMs ([15]) are already designed. However, these designs just enhance the throughput of the interface, not of the memory array.

The capacities of RAMs will further increase by using smaller submicron structures. However, using current technologies this process has a limit due to fundamental scaling problems ([9]). A low power dissipation, a high performance, and a tiny semiconductor structure cannot be achieved together. A solution for this problem may be a multilevel RAM technology. That is, more than two resolvable voltages may be stored in a memory storage cell. Moreover, vertical or even three-dimensional structures may be exploited for storage cells on RAM chip wafers.

The access time for arbitrary addresses may be reduced with the so-called Fast-Cycle RAM architecture ([66]) by Fujitsu and Toshiba. This architecture no longer uses separate row and column addresses but a single address as SRAMs do. Furthermore, pipelining is not only used on column entries but also on rows. That is, the contents of the sense amplifiers are latched into additional row buffers and the next row of the same memory array can be activated during read and write operations in the current one.

Finally, CPU-like functionalities and the RAM array may be coupled more tightly, e.g., intelligent RAM (IRAM [58]), parallel processing RAM (PPRAM [54]), and computing RAM (CRAM [11]) are concepts to integrate RAM with logic circuits. In addition, the coupling of reconfigurable logic and DRAMs is investigated in reconfigurable architecture DRAMs (RADRAM [57]) and was considered in the transit project [6]. Moreover, functional units may operate directly on the contents of the sense amplifiers. Consider that a current DRAM array is able to transfer the contents of a memory row (e.g. 512 Byte) to the sense amplifiers during activation in less than 40 ns. That is, a memory chip can achieve a throughput about 12 GByte per second internally. Most of this potential is currently unused by todays memory interfaces.

References

[1] E. Adler, J.K. DeBrosse, S.F. Geissler, S.J. Holmes, M.D. Jaffe, J.B. Johnson, C.W. Koburger, J.B. Lasky, B. Lloyd, G.L. Miles, J.S. Nakos, W.P. Noble, S.H. Voldman, M. Armacost, and R. Ferguson. The evolution of IBM CMOS DRAM technology. *IBM Journal of Research and Development*, 39(1-2):167–188, March 1995.

[2] Graham Allan. DDR SDRAM/SGRAM, an interpretation of the JEDEC standard. Technical report, Mosaid Technologies Inc., September 1998.

[3] Keith Boland and Apostolos Dollas. Predicting and precluding problems with memory latency. *IEEE Micro*, 14(4):59–67, August 1994.

[4] Doug Burger, James R. Goodman, and Alain Kägi. Memory bandwidth limitations of future microprocessors. *Computer Architecture News*, 24(2):78–89, May 1996.

[5] Richard Crisp. Direct rambus technology: The new main memory standard. *IEEE Micro*, 17(6):18 – 28, November 1997.

[6] Andre DeHon. Notes on integrating reconfigurable logic with DRAM arrays. transit note 120, MIT Transit Project, artifical intelligence laboratory, Massachusetts Institute of Technology, 1995.

[7] Robert H. Dennard. Field-effect transistor memory. *U.S. patent 3 387 286*, June 1968.

[8] Robert H. Dennard. Evolution of the MOSFET dynamic RAM - a personal view. *IEEE Transactions on Electron Devices*, 31(11):1549–1555, November 1984.

[9] Robert H. Dennard. Scaling challenges for DRAM and microprocessors in the 21st century. In *Proceedings of the Sixth International Symposium on Ultralarge Scale Integration Science and Technology. ULSI Science and Technology 1997*, pages 519 – 532. Electrochemical Society, 1997.

[10] Brian Dipert. Advanced DRAM puts you in the fast lane. *EDN-Magazine (US-Edition)*, 42(21), October 1997.

[11] D. Elliott, M. Snelgrove, C. Cojocaru, and M. Stumm. Computing RAMs for media processing. *Proceedings of the SPIE, The International Society for Optical Engineering*, 3021:66–77, 1997.

[12] Enhanced Memory Systems Inc. *16Mbit Enhanced Synchronous DRAM*, February 1998.

[13] Fujitsu Semiconductor. *2 x 256k x 32bit Synchronous Graphic RAM MB81G163222-70.*

[14] Peter Gillingham and Bill Vogley. SLDRAM: High-performance, open-standard memory. *IEEE Micro*, 17(6):29 – 39, November 1997.

[15] Takeshi Hamamoto, Masaki Tsukude, Kazutami Arimoto, Yasuhiro Konishi, Takayuki Miyamoto, Hideyuki Ozaki, and Michihiro Yamada. 400-MHz random column operating SDRAM techniques with self-skew compensation. *IEEE Journal of Solid State Circuits*, 33(5):770 – 778, May 1998.

[16] Hitachi. *256M LVTTL interface SDRAM, HM5225(40,80,16)5A*, rev. 0.1 edition, June 1998.

[17] Hyundai Electronics. *512K x 32bit Synchronous Graphics RAM HY58163210*, November 1997.

[18] Hyundai Semiconductor. *64/72 Mbit Direct RDRAM, HYRDU64164/HYRDU73184*, March 1998.

[19] IBM Corp. *64Mb Direct Rambus DRAM*, November 1997.

[20] IBM Corp. *128Mb Direct Rambus DRAM*, May 1998.

[21] IBM Corp. *16 Mb Enhanced Synchronous DRAM, IBM0516(40/80/16)9CT3A*, November 1998.

[22] IBM Corp. *256Mb Synchronous DRAM, Die Revision A, IBM0325(40,80,16,4B)4*, August 1998.

[23] IBM Corp. *64mb Double Data Rate Synchronous DRAM, IBM0664804ET3A*, August 1998.

[24] IBM Corp. *8Mb (256Kx36 & 512Kx18) and 4Mb (128Kx36 & 256Kx18) SRAM, IBM04(18/36)A(8/4)1QLAA*, November 1998.

[25] IBM Corp. *8Mb (256Kx36 & 512Kx18) and 4Mb (128Kx36 & 256Kx18) SRAM, IBM04(18/36)A(8/4)CXLBB*, November 1998.

[26] IBM Corp. *64Mb Synchronous DRAM, Die Revision C, IBM0364(40,80,16,4B)4*, January 1999.

[27] Intel Corp. *PC SDRAM Registered DIMM Design Support Document*, rev. 1.2 edition, October 1998.

[28] Intel Corp. *PC SDRAM Specification*, rev. 1.63 edition, October 1998.

[29] Intel Corp. *PC SDRAM unbuffered DIMM Specification*, rev. 1.0 edition, February 1998.

[30] Joint Electron Device Engineering Council (JEDEC). *Standard 21C (JESD21C): Configurations for solid state memories: official and preliminary releases.*

[31] Joint Electron Device Engineering Council (JEDEC). *Standard No. 12 Series: CMOS Semicustom Integrated Circuits.*

[32] Joint Electron Device Engineering Council (JEDEC). *Standard No. 8 Series: low voltage interface standards.*

[33] Joint Electron Device Engineering Council (JEDEC) JC-11 committee. *Publication No. 95: Registered and standard outlines for solid state related products, subsection microelectronic outlines.*

[34] Yasunao Katayama. Trends in semiconductor memories. *IEEE Micro*, 17(6):10 – 17, November 1997.

[35] Masaki Kumanoya, Toshiyuki Ogawa, and Kazunari Inoue. Advances in DRAM interfaces. *IEEE Micro*, 15(6):30–36, December 1995.

[36] Masaki Kumanoya, Toshiyuki Ogawa, Yasuhiro Konishi, Katsumi Dosaka, and Kazuhiro Shimotori. Trends in high-speed DRAM architectures. *IEICE Transactions on Electronics*, E79-C(4):472–481, April 1996.

[37] LG Semicon Co.,Ltd. *1MWord x 16 Bit x 4 Bank Synchronous Dynamic RAM, GM72V661641CT/CLT*, rev. 1.0 edition, July 1998.

[38] Micron Technology Inc. *128MEG x4, x8, x16 SDRAM*, March 1998.

[39] Micron Technology Inc. *2.5 V, pipelined, SCD syncburst SRAM 256k x 18, 128k x 32/36*, February 1998.

[40] Micron Technology Inc. *400 Mb/s/pin 4M x 18 SLDRAM, MT49V4M18*, May 1998.

[41] Micron Technology Inc. *4.5 Mb Claymore SRAM, MT57L256H18P, MT57L128H36P*, February 1998.

[42] Micron Technology Inc. *8Mb Syncburst SRAM, pipelined, SCD, 8Mb:512Kx18, 256Kx32/36, MT58L(512/256)L(18/32/36)P*, 1998.

[43] Micron Technology Inc. *Double Data Rate SDRAM, MT46LC8M8-2 Meg x 8 x 4 banks*, May 1998.

[44] Micron Technology Inc. *Double Data Rate SGRAM, 512K x 32 DDR SGRAM, MT45V512K32 - 128K x 32 x 4 banks*, September 1998.

[45] Micron Technology Inc. *LVTTL, Pipelined ZBT SRAM, 256Kx18, 128Kx32/36, MT55L(128/256)L(32/36/18)*, February 1998.

[46] Bruce Millar and Peter Gillingham. Two high-bandwidth memory bus structures. *IEEE Design & Test of Computers*, 16(1):42–52, March 1999.

[47] Mitsubishi Corp. *1048576-bit synchronous burst SRAM M5M5V1132AGP-4*, ver. g edition.

[48] Mitsubishi Corp. *3D-RAM (M5M410092B)*, rev. 0.95 edition.

[49] Mitsubishi Corp. *16M Synchronous Graphics RAM M5M4S16G50DFP-7*, April 1997.

[50] Mitsubishi Corp. *256M Synchronous DRAM M5M4V25S40TP-7*, September 1997.

[51] Mitsubishi Electric. *16M Cached DRAM with 16k SRAM, M5M4V16169DTP/RT-7,-8,-10,-15*, July 1998.

[52] Mosaid Technologies Inc. *4 Bank 256M DDRSDRAM (x4,x8,x16), 256MDDRSDRAM*, July 1998.

[53] MoSys Inc. *MD904 to MD920, 0.5 to 2.5 MByte Multibank DRAM*, February 1997.

[54] K. Murakami, S. Shirakawa, and H. Miyajima. Parallel processing RAM chip with 256 Mb DRAM and quad processors. In *1997 IEEE International Solid-State Circuits Conference. Digest of Technical Papers*, pages 228–229, 1997.

[55] NEC Corp. *64MBit Virtual Channel SDRAM*, October 1998.

[56] Yoichi Oshima, Bing J. Sheu, and Steve H. Jen. High-speed memory architectures for multimedia applications. *IEEE Circuits and Devices Magazine*, 13(1):8–13, January 1997.

[57] Mark Oskin, Frederik T. Chong, and Timothy Sherwood. Active pages: a computation model for intelligent memory. In *Proceedings. 25th Annual International Symposium on Computer Architecture*, pages 192–203. IEEE Computer Society, 1998.

[58] D. Patterson, T. Anderson, N. Cardwell, R. Fromm, K. Keeton, C. Kozyrakis, R. Thomas, and K. Yelick. Intelligent RAM (IRAM): chips that remember and compute. In *1997 IEEE International Solid-State Circuits Conference. Digest of Technical Papers*, pages 224–225, 1997.

[59] Betty Prince. *High Performance Memories.* John Wiley & Sons Ltd., 1996.

[60] Rambus Inc. *Direct RDRAM 64/72-MBit (256K x16/18 x 16d), advance information*, November 1998.

[61] Samsung Electronics. *16Mbit DDR SGRAM, 128K x 32Bit x 4 Banks, Double Data Rate Synchronous Graphic RAM with Bi-directional Data Strobe, KM432D5131*, April 1998.

[62] Samsung Electronics. *16Mbit SGRAM KM4132G512*, March 1998.

[63] Samsung Electronics. *256Kx36 & 512Kx18 Pipelined NtRAM, KM7(36/18)V(8/9)49*, rev. 0.2 edition, September 1998.

[64] Samsung Electronics. *4Mx16Bit x4 Banks Synchronous DRAM, KM416S16230A*, May 1998.

[65] Samsung Electronics. *512K x 32Bit x 4 Banks Synchronous DRAM, KM432S2030B*, July 1998.

[66] Yasuharu Sato, Takaaki Suzuki, Tadao Aikawa, Shin ya Fujioka, Waichiro Fujieda, Hiroyuki Kobayashi, Hitoshi Ikeda, Takayuki Nagasawa, Akihiro Funyu, Yasuhiro Fujii, Ken ichi Kawasaki, Masafumi Yamazaki, and Masao Taguchi. Fast cycle RAM (FCRAM); a 20-ns random row access, pipe-lined operating DRAM. In *1998 Symposium on VLSI Circuits. Digest of Technical Papers*, pages 22–25. IEEE, 1998.

[67] John D. Schmidt. Integrated MOS transistor random access memory. *Solid State Design, Communications & Data Equipment*, 6(1):21–25, January 1965.

[68] Siemens AG. *Multibank DRAMs HYB39M93200, HYB39M83200*, February 1997.

[69] Siemens AG. *16M Synchronous Graphics RAM HYB39S16320TQ-7*, March 1998.

[70] Siemens AG. *256 MBit Synchronous DRAM, HYB39S256400/800/160T*, April 1998.

[71] Siemens Semiconductor Group. *64 MBit Virtual Channel SDRAM, HYB39V64x0yT*, November 1998.

[72] SLDRAM Inc. *400 Mb/s/pin 4M x 18 SLDRAM, pipelined, eight bank, 2.5 V operation, SLD4M18DR400*, July 1998.

[73] Toshiba. *TC59R7218XB*, July 1998.

[74] Toshiba Corp. *524,288 words x 4 banks x 32-bits Synchronous Dynamic RAM, TC59S6432BFT-80*, 1997.

[75] William A. Wulf and Sally A. McKee. Hitting the memory wall: implications of the obvious. *Computer-Architecture-News. vol.23, no.1; March 1995; p.20-4*, 23(1):20 – 24, March 1995.